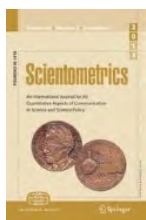


# Exhibit 51

[Home](#) [Scientometrics](#) [Article](#)

# Negative results are disappearing from most disciplines and countries


Published: 11 September 2011 90,891–904 (2012)

[Scientometrics](#)[Aims and scope](#)[Submit manuscript](#)[Daniele Fanelli](#)  12k Accesses  670 Citations  307 Altmetric  41 Mentions[Explore all metrics](#) →[Cite this article](#)

## Abstract

Concerns that the growing competition for funding and citations might distort science are frequently discussed, but have not been verified directly. Of the hypothesized problems, perhaps the most worrying is a worsening of positive–outcome bias. A system that disfavours negative results not only distorts the scientific literature directly, but might also discourage high-risk projects and pressure scientists to fabricate and falsify their data. This study analysed over 4,600 papers published in all disciplines between 1990 and 2007, measuring the frequency of papers that, having declared to have “tested” a hypothesis,

reported a positive support for it. The overall frequency of positive supports has grown by over 22% between 1990 and 2007, with significant differences between disciplines and countries. The increase was stronger in the social and some biomedical disciplines. The United States had published, over the years, significantly fewer positive results than Asian countries (and particularly Japan) but more than European countries (and in particular the United Kingdom). Methodological artefacts cannot explain away these patterns, which support the hypotheses that research is becoming less pioneering and/or that the objectivity with which results are produced and published is decreasing.

 This is a preview of subscription content, [log in via an institution](#)  to check access.

### Access this article

Log in via an institution

Add to cart USD 39.95

Final price calculated at checkout.

Instant access to the full article PDF.

Rent this article via [DeepDyve](#) 

[Institutional subscriptions](#) →

## References

Atkin, P. A. (2002). A paradigm shift in the medical literature. *British Medical Journal*, 325(7378), 1450–1451.

[Article](#) [Google Scholar](#)

Bian, Z. X., & Wu, T. X. (2010). Legislation for trial registration and data transparency. *Trials*, 11, 64. doi:[10.1186/1745-6215-11-64](https://doi.org/10.1186/1745-6215-11-64).

[Article](#) [Google Scholar](#)

Bonitz, M., & Scharnhorst, A. (2001). Competition in science and the Matthew core journals. *Scientometrics*, 51(1), 37–54.

[Article](#) [Google Scholar](#)

Browman, H. I. (1999). The uncertain position, status and impact of negative results in marine ecology: Philosophical and practical considerations. *Marine Ecology Progress Series*, 191, 301–309.

[Article](#) [Google Scholar](#)

Csada, R. D., James, P. C., & Espie, R. H. M. (1996). The “file drawer problem” of non-significant results: Does it apply to biological research? *Oikos*, 76(3), 591–593.

[Article](#) [Google Scholar](#)

de Meis, L., Velloso, A., Lannes, D., Carmo, M. S., & de Meis, C. (2003). The growing competition in Brazilian science: Rites of passage, stress and burnout. *Brazilian Journal of Medical and Biological Research*, 36(9), 1135–1141.

[Article](#) [Google Scholar](#)

De Rond, M., & Miller, A. N. (2005). Publish or perish—Bane or boon of academic life? *Journal of Management Inquiry*, 14(4), 321–329. doi:[10.1177/1056492605276850](https://doi.org/10.1177/1056492605276850).

[Article](#) [Google Scholar](#)



Delong, J. B., & Lang, K. (1992). Are all economic hypotheses false. *Journal of Political Economy*, 100(6), 1257–1272.

[Article](#) [Google Scholar](#)

Doucouliafos, H., Laroche, P., & Stanley, T. D. (2005). Publication bias in union-productivity research? *Relations Industrielles-Industrial Relations*, 60(2), 320–347.

[Google Scholar](#)

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., et al. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE*, 3(8), e3081. [Research Support, Non-U.S. Gov't; Review].

[Article](#) [Google Scholar](#)

Evanschitzky, H., Baumgarth, C., Hubbard, R., & Armstrong, J. S. (2007). Replication research's disturbing trend. *Journal of Business Research*, 60(4), 411–415.  
doi:[10.1016/j.jbusres.2006.12.003](https://doi.org/10.1016/j.jbusres.2006.12.003).

[Article](#) [Google Scholar](#)

Fanelli, D. (2010a). Do pressures to publish increase scientists' bias? An empirical support from US States Data. *Plos One*, 5(4), e10271. doi:[10.1371/journal.pone.0010271](https://doi.org/10.1371/journal.pone.0010271).

[Article](#) [Google Scholar](#)

Fanelli, D. (2010b). “Positive” results increase down the hierarchy of the sciences. *Plos One*, 5(3), e10068. doi:[10.1371/journal.pone.0010068](https://doi.org/10.1371/journal.pone.0010068).

[Article](#) [MathSciNet](#) [Google Scholar](#)

Feigenbaum, S., & Levy, D. M. (1996). Research bias: Some preliminary findings. *Knowledge and Policy: The International Journal of Knowledge Transfer and Utilization*, 9(2 & 3), 135–142.

[Google Scholar](#)

Formann, A. K. (2008). Estimating the proportion of studies missing for meta-analysis due to publication bias. *Contemporary Clinical Trials*, 29(5), 732–739.  
doi:[10.1016/j.cct.2008.05.004](https://doi.org/10.1016/j.cct.2008.05.004).

[Article](#) [Google Scholar](#)

Fronczak, P., Fronczak, A., & Holyst, J. A. (2007). Analysis of scientific productivity using maximum entropy principle and fluctuation-dissipation theorem. *Physical Review E*, 75(2), 026103. doi:[10.1103/PhysRevE.75.026103](https://doi.org/10.1103/PhysRevE.75.026103).

[Article](#) [Google Scholar](#)

Gad-el-Hak, M. (2004). Publish or perish—An ailing enterprise? *Physics Today*, 57(3), 61–62.

[Article](#) [Google Scholar](#)

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research—Do arbitrary significance levels distort published results? *Sociological Methods & Research*, 37(1), 3–30.

[Article](#) [MathSciNet](#) [Google Scholar](#)

Howard, G. S., Hill, T. L., Maxwell, S. E., Baptista, T. M., Farias, M. H., Coelho, C., et al. (2009). What's wrong with research literatures? And how to make them right. *Review of General Psychology*, 13(2), 146–166.

[Article](#) [Google Scholar](#)

Hubbard, R., & Vetter, D. E. (1996). An empirical comparison of published replication research in accounting, economics, finance, management, and marketing. *Journal of Business Research*, 35(2), 153–164.

[Article](#) [Google Scholar](#)

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Plos Medicine*, 2(8), 696–701.

[Article](#) [Google Scholar](#)

Ioannidis, J. P. A. (2006). Evolution and translation of research findings: From to where? *Plos Clinical Trials*, 1, e36. doi:[10.1371/journal.pctr.0010036](https://doi.org/10.1371/journal.pctr.0010036).

[Article](#) [Google Scholar](#)

Ioannidis, J. P. A. (2008a). Perfect study, poor evidence: Interpretation of biases preceding study design. *Seminars in Hematology*, 45(3), 160–166.

[Article](#) [MathSciNet](#) [Google Scholar](#)

Ioannidis, J. P. A. (2008b). Why most discovered true associations are inflated. *Epidemiology*, 19(5), 640–648.

[Article](#) [Google Scholar](#)

Ioannidis, J. P. A., Ntzani, E. E., Trikalinos, T. A., & Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature Genetics*, 29(3), 306–309.

[Article](#) [Google Scholar](#)

Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, 58(6), 543–549.

[Article](#) [Google Scholar](#)

Jeng, M. (2006). A selected history of expectation bias in physics. *American Journal of Physics*, 74(7), 578–583.

[Article](#) [Google Scholar](#)

Jennions, M. D., & Moller, A. P. (2002). Publication bias in ecology and evolution: An empirical assessment using the ‘trim and fill’ method. *Biological Reviews*, 77(2), 211–222.

[Article](#) [Google Scholar](#)

Jennions, M. D., & Moller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology*, 14(3), 438–445.

[Article](#) [Google Scholar](#)

Jones, K. S., Derby, P. L., & Schmidlin, E. A. (2010). An investigation of the prevalence of replication research in human factors. *Human Factors*, 52(5), 586–595.

doi:[10.1177/0018720810384394](https://doi.org/10.1177/0018720810384394).

[Article](#) [Google Scholar](#)

Kelly, C. D. (2006). Replicating empirical research in behavioral ecology: How and why it should be done but rarely ever is. *Quarterly Review of Biology*, 81(3), 221–236.

[Article](#) [Google Scholar](#)

King, D. A. (2004). The scientific impact of nations. *Nature*, 430(6997), 311–316.  
doi:[10.1038/430311a](https://doi.org/10.1038/430311a).

[Article](#) [Google Scholar](#)

Knight, J. (2003). Negative results: Null and void. *Nature*, 422(6932), 554–555.

[Article](#) [Google Scholar](#)

Kundoor, V., & Ahmed, M. K. K. (2010). Uncovering negative results: Introducing an open access journal “Journal of Pharmaceutical Negative Results”. *Pharmacognosy Magazine*, 6(24), 345–347. doi:[10.4103/0973-1296.71783](https://doi.org/10.4103/0973-1296.71783).

[Google Scholar](#)

Kyzas, P. A., Denaxa-Kyza, D., & Ioannidis, J. P. A. (2007). Almost all articles on cancer prognostic markers report statistically significant results. *European Journal of Cancer*, 43(17), 2559–2579.

[Article](#) [Google Scholar](#)

Larsen, P. O., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84(3), 575–603.  
doi:[10.1007/s11192-010-0202-z](https://doi.org/10.1007/s11192-010-0202-z).

[Article](#) [Google Scholar](#)

Lawrence, P. A. (2003). The politics of publication—Authors, reviewers and editors must act to protect the quality of research. *Nature*, 422(6929), 259–261. doi:[10.1038/422259a](https://doi.org/10.1038/422259a).

[Article](#) [Google Scholar](#)

Lortie, C. J. (1999). Over-interpretation: Avoiding the stigma of non-significant results. *Oikos*, 87(1), 183–184.

[Article](#) [Google Scholar](#)

Maddock, J. E., & Rossi, J. S. (2001). Statistical power of articles published in three health psychology-related journals. *Health Psychology*, 20(1), 76–78.

[Article](#) [Google Scholar](#)

Marsh, D. M., & Hanlon, T. J. (2007). Seeing what we want to see: Confirmation bias in animal behavior research. *Ethology*, 113(11), 1089–1098.

[Article](#) [Google Scholar](#)

Meho, L. I. (2007). The rise and rise of citation analysis. *Physics World*, 20(1), 32–36.

[Google Scholar](#)

Nicolini, C., & Nozza, F. (2008). Objective assessment of scientific performances world-wide. *Scientometrics*, 76(3), 527–541. doi:[10.1007/s11192-007-1786-9](https://doi.org/10.1007/s11192-007-1786-9).

[Article](#) [Google Scholar](#)

Osuna, C., Crux-Castro, L., & Sanz-Menedez, L. (2011). Overturning some assumptions about the effects of evaluation systems on publication performance. *Scientometrics*, 86, 575–592.

[Article](#) [Google Scholar](#)

Palmer, A. R. (2000). Quasireplication and the contract of error: Lessons from sex ratios, heritabilities and fluctuating asymmetry. *Annual Review of Ecology and Systematics*, 31, 441–480.

[Article](#) [Google Scholar](#)

Pautasso, M. (2010). Worsening file-drawer problem in the abstracts of natural, medical and social science databases. *Scientometrics*, 85(1), 193–202. doi:[10.1007/s11192-010-0233-5](https://doi.org/10.1007/s11192-010-0233-5).

[Article](#) [Google Scholar](#)

Qiu, J. (2010). Publish or perish in China. *Nature*, 463(7278), 142–143. doi:[10.1038/463142a](https://doi.org/10.1038/463142a).

[Article](#) [Google Scholar](#)

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13(2), 90–100. doi:[10.1037/a0015108](https://doi.org/10.1037/a0015108).

[Article](#) [Google Scholar](#)

Shelton, R. D., Foland, P., & Gorelskyy, R. (2007). Do new SCI journals have a different national bias? *Proceedings of ISSI 2007: 11th international conference of the international society for scientometrics and informetrics, Vols I and II* (pp. 708–717).

Shelton, R. D., Foland, P., & Gorelskyy, R. (2009). Do new SCI journals have a different national bias? *Scientometrics*, 79(2), 351–363. doi:[10.1007/s11192-009-0423-1](https://doi.org/10.1007/s11192-009-0423-1).

[Article](#) [Google Scholar](#)

Silvertown, J., & McConway, K. J. (1997). Does “publication bias” lead to biased science? *Oikos*, 79(1), 167–168.

[Article](#) [Google Scholar](#)

Simera, I., Moher, D., Hirst, A., Hoey, J., Schulz, K. F., & Altman, D. G. (2010). Transparent and accurate reporting increases reliability, utility, and impact of your research: Reporting guidelines and the EQUATOR Network. *Bmc Medicine*, 8, 24. doi:[10.1186/1741-7015-8-24](https://doi.org/10.1186/1741-7015-8-24).

[Article](#) [Google Scholar](#)

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., et al. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment*, 14(8), 1–193. doi:[10.3310/hta14080](https://doi.org/10.3310/hta14080).

[Google Scholar](#)

Statzner, B., & Resh, V. H. (2010). Negative changes in the scientific publication process in ecology: Potential causes and consequences. *Freshwater Biology*, 55(12), 2639–2653. doi:[10.1111/j.1365-2427.2010.02484.x](https://doi.org/10.1111/j.1365-2427.2010.02484.x).

[Article](#) [Google Scholar](#)

Steen, R. G. (2011). Retractions in the scientific literature: Do authors deliberately commit research fraud? *Journal of Medical Ethics*, 37(2), 113–117.

[Article](#) [Google Scholar](#)

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited—The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician*, 49(1), 108–112.



Tsang, E. W. K., & Kwan, K. M. (1999). Replication and theory development in organizational science: A critical realist perspective. *Academy of Management Review*, 24(4), 759–780.

[Google Scholar](#)

Warner, J. (2000). A critical review of the application of citation studies to the Research Assessment Exercises. *Journal of Information Science*, 26(6), 453–459.

[Article](#) [Google Scholar](#)

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydi, O. (2008). Why current publication practices may distort science. *Plos Medicine*, 5(10), 1418–1422. doi:[10.1371/journal.pmed.0050201](https://doi.org/10.1371/journal.pmed.0050201).

[Article](#) [Google Scholar](#)

Yousefi-Nooraie, R., Shakiba, B., & Mortaz-Hejri, S. (2006). Country development and manuscript selection bias: A review of published studies. *BMC Medical Research Methodology*, 6, 37.

[Article](#) [Google Scholar](#)

## Acknowledgments

---

Robin Williams gave helpful comments, and François Briatte crosschecked the coding protocol. This work was supported by a Marie Curie Intra-European Fellowship (Grant Agreement Number PIEF-GA-2008-221441) and a Leverhulme Early-Career fellowship (ECF/2010/0131).

## Author information

---

### Authors and Affiliations

ISSTI-Institute for the Study of Science, Technology and Innovation, The University of Edinburgh, Old Surgeons' Hall, Edinburgh, EH1 1LZ, Scotland, UK

Daniele Fanelli

### Corresponding author

Correspondence to [Daniele Fanelli](#).

### Rights and permissions

---

[Reprints and Permissions](#)

### About this article

---

### Cite this article

Fanelli, D. Negative results are disappearing from most disciplines and countries. *Scientometrics* **90**, 891–904 (2012). <https://doi.org/10.1007/s11192-011-0494-7>

Received

01 July 2011

Published

11 September 2011

Issue Date

March 2012

DOI

<https://doi.org/10.1007/s11192-011-0494-7>

### Keywords

[Bias](#)

[Misconduct](#)

[Research evaluation](#)

[Publication](#)

[Publish or perish](#)

[Competition](#)



# Exhibit 52

## The Logic and Limits of Event Studies in Securities Fraud Litigation

Jill E. Fisch,<sup>\*</sup> Jonah B. Gelbach<sup>\*\*</sup> & Jonathan Klick<sup>\*\*\*</sup>

*Event studies have become increasingly important in securities fraud litigation, and the Supreme Court's 2014 decision in Halliburton Co. v. Erica P. John Fund, Inc. heightened their importance by holding that the results of event studies could be used to obtain or rebut the presumption of reliance at the class certification stage. As a result, getting event studies right has become critical. Unfortunately, courts and litigants widely misunderstand the event study methodology leading, as in Halliburton, to conclusions that differ from the stated standard.*

*This Article provides a primer explaining the event study methodology and identifying the limitations on its use in securities fraud litigation. It begins by describing the basic function of the event study and its foundations in financial economics. The Article goes on to identify special features of securities fraud litigation that cause the statistical properties of event studies to differ from those in the scholarly context in which event studies were developed. Failure to adjust the standard approach to reflect these special features can lead an event study to produce conclusions inconsistent with the standards courts intend to apply. Using the example of the Halliburton litigation, we illustrate the use of these adjustments and demonstrate how they affect the results in that case.*

*The Article goes on to highlight the limitations of event studies and explains how those limitations relate to the legal issues for which they are introduced. These limitations bear upon important normative questions about the role event studies should play in securities fraud litigation.*

---

<sup>\*</sup>Perry Golkin Professor of Law and co-Director, Institute for Law & Economics, University of Pennsylvania Law School.

<sup>\*\*</sup>Professor of Law, University of Pennsylvania Law School.

<sup>\*\*\*</sup>Professor of Law, University of Pennsylvania Law School.

We thank Matthew Adler, Bernard Black, Ryan Bubb, James Cox, Merritt Fox, Jerold Warner, and the many thoughtful participants at Rutgers-Camden, the Penn/NYU Law and Finance Symposium, USC-Gould, Duke, and Boston University for their comments, questions, and suggestions.

INTRODUCTION.....	554
I. THE ROLE OF EVENT STUDIES IN SECURITIES LITIGATION.....	558
II. THE THEORY OF FINANCIAL ECONOMICS AND THE PRACTICE OF EVENT STUDIES: AN OVERVIEW .....	569
A. Steps (1)–(4): Estimating a Security’s Excess Return .....	571
B. Step (5): Statistical Significance Testing in an Event Study ..	573
III. THE EVENT STUDY AS APPLIED TO THE <i>HALLIBURTON</i> LITIGATION ..	579
A. Dates and Events at Issue in the <i>Halliburton</i> Litigation .....	579
B. An Illustrative Event Study of the Six Dates at Issue in the <i>Halliburton</i> Litigation.....	582
IV. SPECIAL FEATURES OF SECURITIES FRAUD LITIGATION AND THEIR IMPLICATIONS FOR THE USE OF EVENT STUDIES .....	588
A. The Inappropriateness of Two-Sided Tests .....	589
B. Non-Normality in Excess Returns .....	593
C. Multiple Event Dates of Interest .....	597
1. <i>When the question of interest is whether any disclosure             had an unusual effect</i> .....	598
2. <i>How should events be grouped together to adjust for             multiple testing?</i> .....	600
3. <i>When the question of interest is whether both of two             event dates had an effect of known sign.</i> .....	602
D. Dynamic Evolution of the Excess Return’s Standard Deviation.....	604
E. Summary and Comparison to the District Court’s Class Certification Order.....	609
V. EVIDENTIARY CHALLENGES TO THE USE OF EVENT STUDIES IN SECURITIES LITIGATION .....	610
A. The Significance of Insignificance .....	611
B. Dealing with Multiple Pieces of News on an Event Date .....	612
C. Power and Type II Error Rates in Event Studies Used in Securities Fraud Litigation.....	615
CONCLUSION .....	617

Introduction

In June 2014, on its second trip to the U.S. Supreme Court, Halliburton scored a partial victory.<sup>1</sup> Halliburton failed to persuade the Supreme Court to overrule its landmark decision in *Basic Inc. v. Levinson*,<sup>2</sup> which had approved the fraud-on-the-market (FOTM) presumption of reliance in private securities fraud litigation.<sup>3</sup> It did, however, persuade the Court to allow defendants to introduce evidence of lack of price impact at class

---

1. Halliburton Co. v. Erica P. John Fund, Inc. (*Halliburton II*), 134 S. Ct. 2398 (2014).  
2. 485 U.S. 224 (1988).  
3. *Halliburton II*, 134 S. Ct. at 2417.

certification.<sup>4</sup> As the Court explained, *Basic* “does not require courts to ignore a defendant’s direct, . . . salient evidence showing that the alleged misrepresentation did not actually affect the stock’s market price and, consequently, that the *Basic* presumption does not apply.”<sup>5</sup>

The concept of price impact<sup>6</sup> is a critical component of securities fraud litigation. Although *Halliburton II* considered price impact only in the context of determining plaintiffs’ reliance on fraudulent statements, price impact is critical to other elements of securities fraud, including loss causation, materiality, and damages. The challenge is how to determine whether fraudulent statements have affected stock price. This task is not trivial—stock prices fluctuate continuously in response to a variety of issuer and market developments as well as “noise” trading. To address the question, litigants use event studies.<sup>7</sup>

Event studies have their origins in the academic literature.<sup>8</sup> Financial economists use event studies to measure the relationship between stock prices and various types of events.<sup>9</sup> The core contribution of the event study is its ability to differentiate between price fluctuations that reflect the range of typical variation for a security and a highly unusual price impact that often may reasonably be inferred from a highly unusual price movement that occurs immediately after an event and has no other potential causes.<sup>10</sup>

4. *Id.*

5. *Id.* at 2416.

6. Fraudulent information has price impact if, in the counterfactual world in which the disclosures were accurate, the price of the security would have been different. One of us has used the related term “price distortion” to encompass both fraudulent information that moves the market price and information that distorts the market by concealing the truth. Jill E. Fisch, *The Trouble with Basic: Price Distortion After Halliburton*, 90 WASH. U. L. REV. 895, 897 n.8 (2013).

7. See, e.g., *In re Oracle Sec. Litig.*, 829 F. Supp. 1176, 1181 (N.D. Cal. 1993) (“Use of an event study or similar analysis is necessary . . . to isolate the influences of [the allegedly fraudulent] information . . .”).

8. See, e.g., *United States v. Schiff*, 602 F.3d 152, 173 n.29 (3d Cir. 2010) (“An event study . . . ‘is a statistical regression analysis that examines the effect of an event [such as the release of information] on a depend[en]t variable, such as a corporation’s stock price.’” (quoting *In re Apollo Group Inc. Sec. Litig.*, 509 F. Supp. 2d 837, 844 (D. Ariz. 2007))).

9. See generally S.P. Kothari & Jerold B. Warner, *Econometrics of Event Studies* (describing the event study literature and conducting census of event studies published in five journals for the years 1974 through 2000), in 1 HANDBOOK OF CORPORATE FINANCE: EMPIRICAL CORPORATE FINANCE 3 (B. Espen Eckbo ed., 2007).

10. See, e.g., Michael J. Kaufman & John M. Wunderlich, *Regressing: The Troubling Dispositive Role of Event Studies in Securities Fraud Litigation*, 15 STAN. J.L. BUS. & FIN. 183, 194 (2009) (citing DAVID TABAK, NERA ECON. CONSULTING, MAKING ASSESSMENTS ABOUT MATERIALITY LESS SUBJECTIVE THROUGH THE USE OF CONTENT ANALYSIS 4 (2007), [http://www.nera.com/content/dam/nera/publications/archive1/PUB\\_Tabak\\_Content\\_Analysis\\_SE\\_C1646-FINAL.pdf](http://www.nera.com/content/dam/nera/publications/archive1/PUB_Tabak_Content_Analysis_SE_C1646-FINAL.pdf) [<https://perma.cc/768L-FPGQ>]) (explaining the role of event studies in identifying an “unusual” price movement).

Use of the event study methodology has become ubiquitous in securities fraud litigation.<sup>11</sup> Indeed, many courts have concluded that the use of an event study is preferred or even required to establish one or more of the necessary elements of the plaintiffs' case.<sup>12</sup> But event studies present challenges in securities fraud litigation. First, it is unclear that courts fully understand event study methodology. For example, Justice Alito asked counsel for the petitioner at oral argument in *Halliburton II*:

Can I ask you a question about these event studies to which you referred? How accurately can they distinguish between . . . the effect on price of the facts contained in a disclosure and an irrational reaction by the market, at least temporarily, to the facts contained in the disclosure?<sup>13</sup>

Counsel responded to Justice Alito's question by stating that: "Event studies are very effective at making that sort of determination."<sup>14</sup> In reality, however, event studies can do no more than demonstrate highly unusual price changes. Event studies do not speak to the rationality of those price changes.

Second, event studies only measure the movement of a stock price in response to the release of unanticipated, material information. In circumstances in which fraudulent statements falsely confirm prior statements, the stock price would not be expected to move.<sup>15</sup> Event studies are not capable of measuring the effect of these so-called confirmatory disclosures on stock price.<sup>16</sup> Similarly, in cases involving multiple "bundled" disclosures, event studies have limited capacity to identify the particular contribution of each piece of information or the degree to which the effects of multiple disclosures may offset each other.<sup>17</sup>

11. See, e.g., Alon Brav & J.B. Heaton, *Event Studies in Securities Litigation: Low Power, Confounding Effects, and Bias*, 93 WASH. U. L. REV. 583, 585 (2015) (observing that "event studies became so entrenched in securities litigation that they are viewed as necessary in every case" (footnotes omitted)).

12. See, e.g., *Bricklayers & Trowel Trades Int'l Pension Fund v. Credit Suisse Sec. (USA) LLC*, 752 F.3d 82, 86 (1st Cir. 2014) ("The usual—it is fair to say 'preferred'—method of proving loss causation in a securities fraud case is through an event study . . .").

13. Transcript of Oral Argument at 24, *Halliburton Co. v. Erica P. John Fund, Inc.* (*Halliburton II*), 134 S. Ct. 2398 (2014) (No. 13-317).

14. *Id.*

15. See, e.g., *Greenberg v. Crossroads Sys., Inc.*, 364 F.3d 657, 665–66 (5th Cir. 2004) ("[C]onfirmatory information has already been digested by the market and will not cause a change in stock price.").

16. As we discuss below, courts have responded to this limitation by allowing plaintiffs to show price impact indirectly through event studies that show a price drop on the date of an alleged corrective disclosure. See, e.g., *In re Vivendi, S.A. Sec. Litig.*, 838 F.3d 223, 259 (2d Cir. 2016) (rejecting "Vivendi's position that an alleged misstatement must be associated with an increase in inflation to have a 'price impact'").

17. This sort of problem, which we discuss below, has arisen in cases; see, e.g., *Archdiocese of Milwaukee Supporting Fund, Inc. v. Halliburton Co.*, No. 3:02–CV–1152–M, 2008 WL 4791492, at \*11 (N.D. Tex. Nov. 4, 2008) (explaining that Halliburton's Dec. 7, 2001 disclosure contained "two distinct components," a corrective disclosure of prior misstatements and new negative



Third, there are important differences between the scholarly contexts for which event studies were originally designed and the use of event studies in securities fraud litigation. For example, academics originally designed the event study methodology to measure the effect of a single event across multiple firms, the effects of multiple events at a single firm, or the effects of multiple events at multiple firms.<sup>18</sup> By contrast, an event study used in securities fraud litigation typically requires evaluating the impact of individual events on a single firm's stock price.<sup>19</sup> These differences have important methodological implications. In addition, determining whether to characterize a price movement as highly unusual is the product of methodological choices, including choices about the level of statistical significance and thus statistical power. In the securities litigation context, those choices have normative implications that courts have not considered.<sup>20</sup> They also may have implications that are inconsistent with governing legal standards.<sup>21</sup>

In this Article, we examine the use of the event study methodology in securities fraud litigation. Part I demonstrates why the concept of a highly unusual price movement is central to a variety of legal issues in securities fraud litigation. Part II explains how event studies work. Part III conducts a stylized event study using data from the *Halliburton* litigation.<sup>22</sup> Part IV identifies the special features of securities fraud litigation that require adjustments to the standard event study approach and demonstrates how a failure to incorporate these features can lead to conclusions inconsistent with

---

information, and denying class certification because plaintiffs were unable to demonstrate that it was more probable than not that the stock price decline was caused by the former); cf. Esther Bruegger & Frederick C. Dunbar, *Estimating Financial Fraud Damages with Response Coefficients*, 35 J. CORP. L. 11, 25 (2009) (explaining that “‘content analysis’ is now part of the tool kit for determining which among a number of simultaneous news events had effects on the stock price”); Alex Rinaudo & Atanu Saha, *An Intraday Event Study Methodology for Determining Loss Causation*, J. FIN. PERSP., July 2014, at 161, 162–63 (explaining how the problem of multiple disclosures can be partially addressed by using an intraday event methodology).

18. See, e.g., Brav & Heaton, *supra* note 11, at 586 (“[A]lmost all academic research event studies are multi-firm event studies (MFESs) that examine large samples of securities from multiple firms.”).

19. See Jonah B. Gelbach, Eric Helland & Jonathan Klick, *Valid Inference in Single-Firm, Single-Event Studies*, 15 AM. L. & ECON. REV. 495, 496–97 (2013) (explaining that securities fraud litigation requires the use of single-firm event studies).

20. See, e.g., *In re Intuitive Surgical Sec. Litig.*, No. 5:13-cv-01920-EJD, 2016 WL 7425926, at \*15 (N.D. Cal. Dec. 22, 2016) (considering plaintiff's argument that “price impact at a 90% confidence level is a statistically significant” effect but ultimately rejecting it because there was “no reason to deviate” from the 95% confidence level adopted by another court).

21. See *infra* Part V.

22. Halliburton announced on December 23, 2016, that it had agreed to a proposed settlement of the case for \$100 million pending court approval. Nate Raymond, *Halliburton Shareholder Class Action to Settle for \$100 Million*, REUTERS (Dec. 23, 2016), <https://www.reuters.com/article/us-halliburton-lawsuit/halliburton-shareholder-class-action-to-settle-for-100-million-idUSKBN14C2BD> [<https://perma.cc/JS9M-DJDD>].

the standards intended by courts. Part V highlights methodological limitations of event studies—i.e., what they can and cannot prove. It also raises questions about whether the 5% significance level typically used in securities litigation is appropriate in light of legal standards of proof. Finally, this Part touches on normative implications that flow from the use of this demanding significance level.

A review of judicial use of event studies raises troubling questions about the capacity of the legal system to incorporate social science methodology, as well as whether there is a mismatch between this methodology and governing legal standards. Our analysis demonstrates that the proper use of event studies in securities fraud litigation requires care, both in a better understanding of the event study methodology and in an appreciation of its limits.

## I. The Role of Event Studies in Securities Litigation

In this Part, we take a systematic look at the different questions that event studies might answer in a securities fraud case.<sup>23</sup> As noted above, the use of event studies in securities fraud litigation is widespread. As litigants and courts have become familiar with the methodology, they have used event studies to address a variety of legal issues.

The Supreme Court's decision in *Basic Inc. v. Levinson* marked the starting point. In *Basic*, the Court accepted the FOTM presumption which holds that "the market price of shares traded on well-developed markets reflects all publicly available information, and, hence, any material misrepresentations."<sup>24</sup> The Court observed that the typical investor, in "buy[ing] or sell[ing] stock at the price set by the market[,] does so in reliance on the integrity of that price."<sup>25</sup> As a result, the Court concluded that an investor's reliance could be presumed for purposes of a 10b-5 claim if the following requirements were met: (i) the misrepresentations were publicly known; (ii) "the misrepresentations were material"; (iii) the stock was "traded [i]n an efficient market"; and (iv) "the plaintiff traded . . . between the time the misrepresentations were made and . . . [when] the truth was revealed."<sup>26</sup>

---

23. To succeed on a federal securities fraud claim, the plaintiff must establish the following elements: "(1) a material misrepresentation (or omission); (2) scienter, i.e., a wrongful state of mind; (3) a connection with the purchase or sale of a security; (4) reliance . . . ; (5) economic loss; and (6) 'loss causation,' i.e., a causal connection between the material misrepresentation and the loss." *Dura Pharm., Inc. v. Broudo*, 544 U.S. 336, 341–42 (2005) (cleaned up).

24. *Basic Inc. v. Levinson*, 485 U.S. 224, 246 (1988).

25. *Id.* at 247.

26. *Id.* at 248 n.27.

The Court's decision in *Basic* was influenced by a law review article by Professor Daniel Fischel of the University of Chicago Law School.<sup>27</sup> Fischel argued that FOTM offered a more coherent approach to securities fraud than then-existing practice because it recognized the market model of the investment decision.<sup>28</sup> Although *Basic* focused on the reliance requirement, Fischel argued that the only relevant inquiry in a securities fraud case was the extent to which market prices were distorted by fraudulent information—it was unnecessary for the court to make separate inquiries into materiality, reliance, causation, and damages.<sup>29</sup> Moreover, Fischel stated that the effect of fraudulent conduct on market price could be determined through a blend of financial economics and applied statistics. Although Fischel did not use the term “event study” in this article, he described the event study methodology.<sup>30</sup>

The lower courts initially responded to the *Basic* decision by focusing extensively on the efficiency of the market in which the securities traded.<sup>31</sup> The leading case on market efficiency, *Cammer v. Bloom*,<sup>32</sup> involved a five-factor test:

(1) the stock's average weekly trading volume; (2) the number of securities analysts that followed and reported on the stock; (3) the presence of market makers and arbitrageurs; (4) the company's eligibility to file a Form S-3 Registration Statement; and (5) a cause-and-effect relationship, over time, between unexpected corporate events or financial releases and an immediate response in stock price.<sup>33</sup>

Economists serving as expert witnesses generally use event studies to address the fifth *Cammer* factor.<sup>34</sup> In this context, the event study is used to determine the extent to which the market for a particular stock responds to new information. Experts generally look at multiple information or news

27. Daniel R. Fischel, *Use of Modern Finance Theory in Securities Fraud Cases Involving Actively Traded Securities*, 38 BUS. LAW. 1 (1982).

28. *Id.* at 2, 9–10.

29. *Id.* at 13.

30. *Id.* at 17–18.

31. See Fisch, *supra* note 6, at 911 (explaining how, after *Basic*, the majority of challenges to class certification involved challenges of “the efficiency of the market in which the securities traded”).

32. 711 F. Supp. 1264 (D.N.J. 1989).

33. DAVID TABAK, NERA ECON. CONSULTING, DO COURTS COUNT *CAMMER* FACTORS? 2 (2012) (quoting *In re Xcelera.com Sec. Litig.*, 430 F.3d 503, 511 (1st Cir. 2005)), [http://www.nera.com/content/dam/nera/publications/archive2/PUB\\_Cammer\\_Factors\\_0812.pdf](http://www.nera.com/content/dam/nera/publications/archive2/PUB_Cammer_Factors_0812.pdf) [<https://perma.cc/75TK-4B4Z>].

34. See *Teamsters Local 445 Freight Div. Pension, Fund v. Bombardier Inc.*, 546 F.3d 196, 207 (2d Cir. 2008) (explaining that the fifth *Cammer* factor—which requires evidence tending to demonstrate that unexpected corporate events or financial releases cause an immediate response in the price of a security—is the most important indicator of market efficiency). But see TABAK, *supra* note 33, at 2–3 (providing evidence that courts are simply “counting” the *Cammer* factors).

events—some relevant to the litigation in question and some not—and evaluate the extent to which these events are associated with price changes in the expected directions.<sup>35</sup>

A number of commentators have questioned the centrality of market efficiency to the *Basic* presumption, disputing either the extent to which the market is as efficient as presumed by the *Basic* court<sup>36</sup> or the relevance of market efficiency altogether.<sup>37</sup> Financial economists do not consider the *Cammer* factors to be reliable for purposes of establishing market efficiency in academic research.<sup>38</sup> Nonetheless, it has become common practice for both plaintiffs and defendants to submit event studies that address the extent to which the market price of the securities in question respond to publicly reported events for the purpose of addressing *Basic*'s requirement that the securities were traded in an efficient market.<sup>39</sup>

*Basic* signaled a broader potential role for event studies, however. By focusing on the harm resulting from a misrepresentation's effect on stock price rather than on the autonomy of investors' trading decisions, *Basic* distanced federal securities litigation from the individualized tort of common law fraud.<sup>40</sup> In this sense, *Basic* was transformative—it introduced a market-based approach to federal securities fraud litigation.<sup>41</sup> Price impact is a critical component of this approach because absent an impact on stock price, plaintiffs who trade in reliance on the market price are not defrauded. As the Supreme Court subsequently noted in *Halliburton II*, “[i]n the absence of

35. See, e.g., *Halliburton Co. v. Erica P. John Fund, Inc. (Halliburton II)*, 134 S. Ct. 2398, 2415 (2014) (“EPJ Fund submitted an event study of various episodes that might have been expected to affect the price of Halliburton’s stock, in order to demonstrate that the market for that stock takes account of material, public information about the company.”).

36. See, e.g., Jonathan R. Macey et al., *Lessons from Financial Economics: Materiality, Reliance, and Extending the Reach of Basic v. Levinson*, 77 VA. L. REV. 1017, 1018 (1991) (citing “substantial disagreement . . . about to what degree markets are efficient, how to test for efficiency, and even the definition of efficiency”). See also Baruch Lev & Meiring de Villiers, *Stock Price Crashes and 10b-5 Damages: A Legal, Economic, and Policy Analysis*, 47 STAN. L. REV. 7, 20 (1994) (“[O]verwhelming empirical evidence suggests that capital markets are not fundamentally efficient.”). Notably, Lev and de Villiers concede that markets are likely information-efficient, which is the predicate requirement for FOTM. See *id.* at 21 (“While capital markets are in all likelihood not fundamentally efficient, widely held and heavily traded securities are probably ‘informationally efficient.’”).

37. Fisch, *supra* note 6, at 898 (“[M]arket efficiency is neither a necessary nor a sufficient condition to establish that misinformation has distorted prices . . .”); see, e.g., Brief of Law Professors as Amici Curiae in Support of Petitioners at 4–5, *Halliburton Co. v. Erica P. John Fund, Inc. (Halliburton II)*, 134 S. Ct. 2398 (2014) (No. 13-317) (arguing that inquiry into market efficiency to show reliance was “unnecessary and counterproductive”).

38. Brav & Heaton, *supra* note 11, at 601.

39. See *Halliburton II*, 134 S. Ct. at 2415 (explaining that both plaintiffs and defendants introduce event studies at the class certification stage for the purpose of addressing market efficiency).

40. See generally Fisch, *supra* note 6, at 913–14.

41. *Id.* at 916.

price impact, *Basic*'s fraud-on-the-market theory and presumption of reliance collapse."<sup>42</sup>

The importance of price impact extends beyond the reliance requirement. In *Dura Pharmaceuticals*,<sup>43</sup> the plaintiffs, relying on *Basic*, filed a complaint in which they alleged that at the time they purchased Dura stock, its price had been artificially inflated due to Dura's alleged misstatements.<sup>44</sup> The Supreme Court reasoned that while artificial price inflation at the time of the plaintiffs' purchase might address the reliance requirement, plaintiffs were also required to plead and prove the separate element of loss causation.<sup>45</sup> Key to the Court's reasoning was that purchasing at an artificially inflated price did not automatically cause economic harm because an investor might purchase at an artificially inflated price and subsequently sell while the price was still inflated.<sup>46</sup>

Following *Dura*, courts allowed plaintiffs to establish loss causation in various ways, but the standard approach involved the use of an event study "to demonstrate both that the economic loss occurred and that this loss was proximately caused by the defendant's misrepresentation."<sup>47</sup> Practically speaking, plaintiffs in the post-*Dura* era need to plead price impact both at the time of the misrepresentation<sup>48</sup> and on the alleged corrective disclosure date. However, in *Halliburton I*,<sup>49</sup> the Supreme Court explained that plaintiffs do not need to prove loss causation to avail themselves of the *Basic* presumption since this presumption has to do with "transaction causation"—the decision to buy the stock in the first place, which occurs before any evidence of loss causation could exist.<sup>50</sup>

42. *Halliburton II*, 134 S. Ct. at 2414.

43. *Dura Pharm., Inc. v. Broudo*, 544 U.S. 336 (2005).

44. *Id.* at 339–40.

45. *Id.* at 346. The Private Securities Litigation Reform Act (PSLRA) codified the loss causation requirement that had previously been developed by lower courts. 15 U.S.C. § 78u-4(b)(4) (1995); see Jill E. Fisch, *Cause for Concern: Causation and Federal Securities Fraud*, 94 IOWA L. REV. 811, 813 (2009) (describing judicial development of the loss causation requirement).

46. *Dura*, 544 U.S. at 342–43.

47. Kaufman & Wunderlich, *supra* note 10, at 198.

48. The former requirement is not necessary in cases involving confirmatory disclosures. See *infra* notes 75–86 and accompanying text (discussing confirmatory disclosures).

49. *Erica P. John Fund, Inc. v. Halliburton Co. (Halliburton I)*, 563 U.S. 804 (2011).

50. *Id.* at 812. As to the merits, though, plaintiffs must also demonstrate a causal link between the two events—the initial misstatement and the corrective disclosure. See, e.g., *Aranaz v. Catalyst Pharm. Partners Inc.*, 302 F.R.D. 657, 671–72 (S.D. Fla. 2014) (describing and rejecting defendants' argument that other information on the date of the alleged corrective disclosure was responsible for the fall in stock price). *Halliburton I* was spawned because the district court had denied class certification on the ground that plaintiffs had failed to persuade the court that there was such a causal link (even though plaintiffs had presented an event study showing a price impact from the misstatements). *Archdiocese of Milwaukee Supporting Fund, Inc. v. Halliburton Co.*, No. 3:02–CV–1152–M, 2008 WL 4791492, at \*1 (N.D. Tex. Nov. 4, 2008).

Plaintiffs responded to *Dura*'s loss causation requirement by presenting event studies showing that the stock price declined in response to an issuer's corrective disclosure. As the First Circuit recently explained: "The usual—it is fair to say 'preferred'—method of proving loss causation in a securities fraud case is through an event study . . . ."<sup>51</sup>

Proof of price impact for purposes of analyzing reliance and causation also overlaps with the materiality requirement.<sup>52</sup> The Court has defined material information as information that has a substantial likelihood to be "viewed by the reasonable investor as having significantly altered the 'total mix' of information made available."<sup>53</sup> Because market prices are a reflection of investors' trading decisions, information that is relevant to those trading decisions has the capacity to impact stock prices, and similarly, information that does not affect stock prices is arguably immaterial.<sup>54</sup> As the Third Circuit explained in *Burlington Coat Factory*:<sup>55</sup> "In the context of an 'efficient' market, the concept of materiality translates into information that alters the price of the firm's stock."<sup>56</sup> Event studies can be used to demonstrate the impact of fraudulent statements on stock price, providing evidence that the statements are material.<sup>57</sup> The lower courts have, on occasion, accepted the argument that the absence of price impact demonstrates the immateriality of alleged misrepresentations.<sup>58</sup>

51. *Bricklayers & Trowel Trades Int'l Pension Fund v. Credit Suisse Sec. (USA) LLC*, 752 F.3d 82, 86 (1st Cir. 2014).

52. *See, e.g., Erica P. John Fund, Inc. v. Halliburton Co.*, 718 F.3d 423, 434–35 n.10 (5th Cir. 2013) ("[T]here is a fuzzy line between price impact evidence directed at materiality and price impact evidence broadly directed at reliance.").

53. *Basic Inc. v. Levinson*, 485 U.S. 224, 231–32 (1988) (quoting *TSC Indus., Inc. v. Northway, Inc.*, 426 U.S. 438, 449 (1976)).

54. *See* Fredrick C. Dunbar & Dana Heller, *Fraud on the Market Meets Behavioral Finance*, 31 DEL. J. CORP. L. 455, 509 (2006) ("The definition of immaterial information . . . is that it is already known or . . . does not have a statistically significant effect on stock price in an efficient market."). *But cf.* Donald C. Langevoort, *Basic at Twenty: Rethinking Fraud on the Market*, 2009 WIS. L. REV. 151, 173–77 (2009) (arguing that in some cases material information may not affect stock prices).

55. *In re Burlington Coat Factory Sec. Litig.*, 114 F.3d 1410 (3d Cir. 1997).

56. *Id.* at 1425.

57. *See, e.g., In re Sadia, S.A. Sec. Litig.*, 269 F.R.D. 298, 302, 311 & n.104, 316 (S.D.N.Y. 2010) (finding that the plaintiffs offered sufficient evidence—among which was an event study conducted by an expert witness—to conclude that the defendant's misstatements were material); *In re Gaming Lottery Sec. Litig.*, No. 96 Civ. 5567(RPP), 2000 WL 193125, at \*1 (S.D.N.Y. Feb. 16, 2000) (describing the event study as "an accepted method for the evaluation of materiality damages to a class of stockholders in a defendant corporation").

58. *See In re Merck & Co. Sec. Litig.*, 432 F.3d 261, 269, 273–75 (3d Cir. 2005) (holding that a false disclosure is immaterial when there is "no negative effect" on a company's stock price directly following the disclosure's publication); *Oran v. Stafford*, 226 F.3d 275, 282 (3d Cir. 2000) (Alito, J.) ("[I]n an efficient market 'the concept of materiality translates into information that alters the price of the firm's stock' . . . ." (quoting *In re Burlington Coat Factory*, 114 F.3d at 1425)).



A statement can be immaterial because it is unimportant or because it conveys information that is already known to the market.<sup>59</sup> The latter argument is known as the “truth on the market” defense since the argument is that the market already knew the truth. According to the truth-on-the-market defense, an alleged misrepresentation that occurs after the market already knows the truth cannot change market perceptions of firm value because any effect of the truth will already have been incorporated into the market price.<sup>60</sup>

In *Amgen*,<sup>61</sup> the parties agreed that the market for Amgen’s stock was efficient and that the statements in question were public, but they disputed the reasons why Amgen’s stock price had dropped on the alleged corrective disclosure dates.<sup>62</sup> Specifically, the defendants argued that because the truth regarding the alleged misrepresentations was publicly known before plaintiffs purchased their shares, plaintiffs did not trade at a price that was impacted by the fraud.<sup>63</sup> Although the majority in *Amgen* concluded that proof of materiality was not required at the class certification stage, it acknowledged that the defendant’s proffered truth-on-the-market evidence could potentially refute materiality.<sup>64</sup>

Proof of economic loss and damages also overlaps proof of loss causation. For plaintiffs to recover damages, they must show that they suffered an economic loss that was caused by the alleged fraud.<sup>65</sup> The 1934 Act provides that plaintiffs may recover actual damages, which must be

59. See *Conn. Ret. Plans & Trust Funds v. Amgen Inc.*, 660 F.3d 1170, 1177 (9th Cir. 2011) (“[T]he truth-on-the-market defense is a method of refuting an alleged misrepresentation’s materiality.” (emphasis omitted)).

60. See, e.g., *Aranaz v. Catalyst Pharm. Partners Inc.*, 302 F.R.D. 657, 670–71 (S.D. Fla. 2014) (explaining that the defendants sought to show that because the market already “knew the truth,” the price was not distorted by alleged misrepresentations).

61. *Amgen Inc. v. Conn. Ret. Plans & Trust Funds*, 568 U.S. 455 (2013).

62. *Id.* at 459, 464; see also Memorandum of Points and Authorities in Opposition to Lead Plaintiff’s Motion for Class Certification at 23, *Conn. Ret. Plans & Trust Funds v. Amgen, Inc.*, No. CV 07-2536 PSG (PLAx), 2009 WL 2633743 (C.D. Cal. Aug. 12, 2009):

Defendants have made a ‘showing’ both that information was publicly available *and* that the market drops that Plaintiff relies on to establish loss causation were not caused by the revelation of any allegedly concealed information. . . . Rather, as Defendants have shown, the market was ‘privity’ to the truth, and the price drops were the result of third-parties’ reactions to public information.

63. *Amgen*, 568 U.S. at 459, 464. As a lower court had put it, “FDA announcements and analyst reports about Amgen’s business [had previously] publicized the truth about the safety issues looming over Amgen’s drugs . . .” *Conn. Ret. Plans & Trust Funds*, 660 F.3d at 1177.

64. See *Amgen*, 568 U.S. at 481–82 (concluding that truth-on-the-market evidence is a matter for trial or for a summary judgment motion, not for determining class certification).

65. 15 U.S.C. § 78u-4(b)(4) (2010). This provision places the burden of establishing loss causation on the plaintiffs in any private securities fraud action brought under Chapter 2B of Title 15. See *Dura Pharm., Inc. v. Broudo*, 544 U.S. 336, 338 (2005) (“A private plaintiff who claims securities fraud must prove that the defendant’s fraud caused an economic loss.” (citing § 78u-4(b)(4))).

proved.<sup>66</sup> A plaintiff who can prove damages has obviously proved she sustained an economic loss. At the same time, a plaintiff who cannot prove damages cannot prove she suffered an economic loss. Thus the economic loss and damages elements merge into one. A number of courts have rejected testimony or reports by damages experts that failed to include an event study.<sup>67</sup>

Notably, while the price impact at the time of the fraud (required in order to obtain the *Basic* presumption of reliance) is not the same as price impact at the time of the corrective disclosures (loss causation under *Dura*),<sup>68</sup> in many cases, the parties may seek to address both elements with a single event study. This is most common in cases that involve alleged fraudulent confirmatory statements. Misrepresentations that falsely confirm market expectations will not lead to an *observable change in price*.<sup>69</sup> But this does not mean they have no *price impact*. As the Second Circuit explained in *Vivendi*,<sup>70</sup> “a statement may cause inflation not simply by *adding* it to a stock, but by *maintaining* it.”<sup>71</sup> The relevant price impact is simply counterfactual: the price would have fallen had there not been fraud.<sup>72</sup>

In cases where plaintiffs allege confirmatory misrepresentations, event study evidence has no probative value related to the alleged misrepresentation dates since the plaintiffs’ own allegations predict no change in price. Thus there will be no *observed* price impact on alleged misrepresentation dates. However, a change in observed price will ultimately occur when the fraud is revealed via corrective disclosures. That is why it is

66. 15 U.S.C. § 78bb(a)(1) (2012).

67. See, e.g., *In re Imperial Credit Indus., Inc. Sec. Litig.*, 252 F. Supp. 2d 1005, 1015 (C.D. Cal. 2003) (“Because of the need ‘to distinguish between the fraud-related and non-fraud related influences of the stock’s price behavior,’ a number of courts have rejected or refused to admit into evidence damages reports or testimony by damages experts in securities cases which fail to include event studies or something similar.” (quoting *In re Oracle Sec. Litig.*, 829 F. Supp. 1176, 1181 (N.D. Cal. 1993))); *In re N. Telecom Ltd. Sec. Litig.*, 116 F. Supp. 2d 446, 460 (S.D.N.Y. 2000) (terming expert’s testimony “fatally deficient in that he did not perform an event study or similar analysis”); *In re Exec. Telecard, Ltd. Sec. Litig.*, 979 F. Supp. 1021, 1025 (S.D.N.Y. 1997) (“The reliability of the Expert Witness’ proposed testimony is called into question by his failure to indicate . . . whether he conducted an ‘event study’ . . .”).

68. See *Erica P. John Fund, Inc. v. Halliburton Co. (Halliburton I)*, 563 U.S. 804, 805 (2011) (distinguishing between reliance and loss causation); see also Fisch, *supra* note 6, at 899 & n.20 (highlighting the distinction and terming the former *ex ante* price distortion and the latter *ex post* price distortion).

69. See, e.g., *FindWhat Inv’r Grp. v. FindWhat.com*, 658 F.3d 1282, 1310 (11th Cir. 2011) (“A corollary of the efficient market hypothesis is that disclosure of confirmatory information—or information already known by the market—will not cause a change in the stock price. This is so because the market has already digested that information and incorporated it into the price.”).

70. *In re Vivendi, S.A. Sec. Litig.*, 838 F.3d 223 (2d Cir. 2016).

71. *Id.* at 258.

72. The *Vivendi* court explained that “once a company chooses to speak, the proper question for purposes of our inquiry into price impact is not what might have happened had a company remained silent, but what would have happened if it had spoken *truthfully*.” *Id.*



appropriate to allow plaintiffs to use event studies concerning dates of alleged corrective disclosures to establish price impact for cases involving confirmatory alleged misrepresentations. A showing that the stock price responded to a subsequent corrective disclosure can provide indirect evidence of the counterfactual price impact of the alleged misrepresentation.<sup>73</sup> Such a conclusion opens the door to consideration of the type of event study conducted for purposes of loss causation, as we discuss below.<sup>74</sup>

*Halliburton II* presented this scenario. Plaintiffs alleged that Halliburton made a variety of fraudulent confirmatory disclosures that artificially maintained the company's stock price.<sup>75</sup> Initially, defendants had argued that the plaintiff could not establish loss causation because Halliburton's subsequent corrective disclosures did not impact the stock price.<sup>76</sup> When the Supreme Court held in *Halliburton I* that the plaintiffs were not required to prove loss causation on a motion for class certification,<sup>77</sup> "Halliburton argued on remand that the evidence it had presented to disprove loss causation also demonstrated that none of the alleged misrepresentations actually impacted Halliburton's stock price, i.e., there was a lack of 'price impact,' and, therefore, Halliburton had rebutted the *Basic* presumption."<sup>78</sup> Halliburton attempted to present "extensive evidence of no price impact," evidence that the lower courts ruled was "not appropriately considered at class certification."<sup>79</sup>

The Supreme Court disagreed. In *Halliburton II*, Chief Justice Roberts explained that the Court's decision was not a bright-line choice between allowing district courts to consider price impact evidence at class certification or requiring them to consider the issue at a later point in trial; price impact evidence from event studies was often already before the court at the class certification stage because plaintiffs were using event studies to demonstrate market efficiency, and defendants were using event studies to counter this

---

73. See *IBEW Local 98 Pension Fund v. Best Buy Co.*, 818 F.3d 775, 782 (8th Cir. 2016) (noting the lower court's reasoning that price impact can be shown when a revelation of fraud is followed by a decrease in price); *In re Bank of Am. Corp. Sec., Derivative, & Emp. Ret. Income Sec. Act (ERISA) Litig.*, 281 F.R.D. 134, 143 (S.D.N.Y. 2012) (finding that stock price's negative reaction to corrective disclosure served to defeat defendant's argument on lack of price impact).

74. See *infra* text accompanying notes 80–89.

75. *Halliburton Co. v. Erica P. John Fund, Inc. (Halliburton II)*, 134 S. Ct. 2398, 2405–06 (2014).

76. Defendant Halliburton Co.'s Brief in Support of the Motion to Dismiss Plaintiffs' Fourth Consol. Class Action Complaint at 22, *Archdiocese of Milwaukee Supporting Fund, Inc. v. Halliburton Co.*, No. 3:02–CV–1152–M, 2008 WL 4791492 (N.D. Tex. Nov. 4, 2008).

77. *Erica P. John Fund, Inc. v. Halliburton Co. (Halliburton I)*, 563 U.S. 804, 813 (2011).

78. *Erica P. John Fund, Inc. v. Halliburton Co.*, 309 F.R.D. 251, 255–56 (N.D. Tex. 2015).

79. *Erica P. John Fund, Inc. v. Halliburton Co.*, 718 F.3d 423, 435 n.11 (5th Cir. 2013), *vacated*, 134 S. Ct. 2398 (2014).

evidence.<sup>80</sup> Under these circumstances, the Chief Justice concluded that prohibiting a court from relying on this same evidence to evaluate whether the fraud affected stock price “makes no sense.”<sup>81</sup>

Because the question of price impact itself is unavoidably before the Court upon a motion for class certification, the Chief Justice explained that the Court’s actual choice concerned merely the *type* of evidence it would allow parties to use in demonstrating price impact on the dates of alleged misrepresentations or alleged corrective disclosures. “The choice . . . is between limiting the price impact inquiry before class certification to indirect evidence”—evidence directed at establishing market efficiency in general—“or allowing consideration of direct evidence as well.”<sup>82</sup> The direct evidence the Court’s majority determined to allow—concerning price impact on dates of alleged misrepresentations and alleged corrective disclosures—will typically be provided in the form of event studies.

On remand, the trial court considered the event study submitted by Halliburton’s expert, which purported to find that neither the alleged misrepresentations nor the corrective disclosures<sup>83</sup> identified by the plaintiff impacted Halliburton’s stock price.<sup>84</sup> After carefully considering the event studies submitted by both parties, which addressed six corrective disclosures, the court found that Halliburton had successfully demonstrated a lack of price impact as to five of the dates and granted class certification with respect to the December 7 alleged corrective disclosure.<sup>85</sup> For several dates, this conclusion was based on the district court’s determination that the event effects were statistically insignificant at the 5% significance level (equivalently, at the 95% confidence level).<sup>86</sup>

Following *Halliburton II*, several other lower courts have considered defendants’ use of event studies to demonstrate the absence of price impact. In *Local 703, I.B. of T. Grocery v. Regions Financial Corp.*,<sup>87</sup> the court of appeals concluded that the defendant had provided evidence that the stock

80. *Halliburton Co. v. Erica P. John Fund, Inc. (Halliburton II)*, 134 S. Ct. 2398, 2417 (2014). The *Halliburton* litigation provides an odd context in which to make this determination since Halliburton had not disputed the efficiency of the public market in its stock. *Archdiocese of Milwaukee Supporting Fund, Inc.*, 2008 WL 4791492, at \*1.

81. *Halliburton II*, 134 S. Ct. at 2415.

82. *Id.* at 2417.

83. As the court explained: “Measuring price change at the time of the corrective disclosure, rather than at the time of the corresponding misrepresentation, allows for the fact that many alleged misrepresentations conceal a truth.” *Halliburton Co.*, 309 F.R.D. at 262.

84. *Id.* at 262–63. The court noted that the expert attributed the one date on which the stock experienced a highly unusual price movement as a reaction to factors other than Halliburton’s disclosure. *Id.*

85. *Id.* at 280.

86. *Id.* at 270.

87. *Local 703, I.B. of T. Grocery & Food Emps. Welfare Fund v. Regions Fin. Corp.*, 762 F.3d 1248 (11th Cir. 2014).

price did not change in light of the misrepresentations and that the trial court, acting prior to *Halliburton II*, “did not fully consider this evidence.”<sup>88</sup> Accordingly, the court vacated and “remand[ed] for fuller consideration . . . of all the price-impact evidence submitted below.”<sup>89</sup> On remand, defendants argued that they had successfully rebutted the *Basic* presumption by providing evidence of no price impact on both the misrepresentation date and the date of the corrective disclosure.<sup>90</sup> The trial court disagreed. The court reasoned that the defendants’ own expert conceded that the 24% decline in the issuer’s stock on the date of the corrective disclosure was far greater than the New York Stock Exchange’s 6.1% decline that day and that given this discrepancy the defense had not shown the absence of price impact.<sup>91</sup> This decision places the burden of persuasion concerning price impact squarely on the defendants.<sup>92</sup>

In *Aranaz v. Catalyst Pharmaceutical Partners Inc.*,<sup>93</sup> the district court permitted the defendant an opportunity to rebut price impact at class certification.<sup>94</sup> The *Aranaz* court explained, however, that the defendant was limited to direct evidence that the alleged misrepresentations had no impact on stock price.<sup>95</sup> The defendants conceded that the stock price rose by 42% on the date of the allegedly misleading press release and fell by 42% on the date of the corrective disclosure<sup>96</sup> but argued that other statements in the two publications caused the “drastic changes in stock price.”<sup>97</sup> The court

88. *Id.* at 1258.

89. *Id.* at 1258–59.

90. Local 703, I.B. of T. Grocery & Food Emps. Welfare Fund v. Regions Fin. Corp., No. CV–10–J–2847–S, 2014 WL 6661918, at \*5–9 (N.D. Ala. Nov. 19, 2014).

91. *Id.* at \*8–10. Defendants argued that their expert’s event study “conclusively finds no price impact on January 20, 2009,” the date of the alleged disclosure. *Id.* at \*8.

92. See Merritt B. Fox, *Halliburton II: It All Depends on What Defendants Need to Show to Establish No Impact on Price*, 70 BUS. LAW. 437, 449, 463 (2015) (describing the resulting statistical burden this approach would impose on defendants to rebut the presumption).

93. 302 F.R.D. 657 (S.D. Fla. 2014).

94. *Id.* at 669–73.

95. *Id.* at 670 (citing *Amgen Inc. v. Conn. Ret. Plans & Trust Funds*, 133 S. Ct. 1184, 1197 (2013)). Under *Halliburton I* and *Amgen*, this limit is appropriate. The district court in *Halliburton* took the same approach on remand following *Halliburton II*. See *Erica P. John Fund, Inc. v. Halliburton Co.*, 309 F.R.D. 251, 261–62 (N.D. Tex. 2015) (“This Court holds that *Amgen* and *Halliburton I* strongly suggest that the issue of whether disclosures are [actually] corrective is not a proper inquiry at the certification stage. *Basic* presupposes that a *misrepresentation* is reflected in the market price at the time of the transaction.” (citing *Halliburton Co. v. Erica P. John Fund, Inc. (Halliburton II)*, 134 S. Ct. 2398, 2416 (2014))). And “at this stage of the proceedings, the Court concludes that the asserted misrepresentations were, in fact, misrepresentations, and assumes that the asserted corrective disclosures were corrective of the alleged misrepresentations.” The court continued to explain that “[w]hile it may be true that a finding that a particular disclosure was not corrective as a matter of law would” break “‘the link between the alleged misrepresentation and . . . the price received (or paid) by the plaintiff . . .,’ the Court is unable to unravel such a finding from the materiality inquiry.” (quoting *Halliburton II*, 134 S. Ct. at 2415–16)).

96. *Aranaz*, 302 F.R.D. at 669.

97. *Id.* at 671.

concluded that because the defendant had the burden of proving that “price impact is *inconsistent* with the results of their analysis,”<sup>98</sup> their evidence was not sufficient to show an absence of price impact. This determination as to the burden of persuasion tracks the approach taken by the *Local 703* court discussed above. Further, following *Amgen*, the *Aranaz* court ruled that the truth-on-the-market defense would not defeat class certification because it concerns materiality and not price impact.<sup>99</sup>

The lower court decisions following *Halliburton II* demonstrate the growing importance of event studies. The most recent trial court decision as to class certification in the *Halliburton* litigation itself<sup>100</sup> demonstrates as well the challenges for the court in evaluating the event study methodology, an issue we will consider in more detail in Part III below.

Significantly, as reflected in the preceding discussion, proof of price impact is relevant to multiple elements of securities fraud. A single event study may provide evidence relating to materiality, reliance, loss causation, economic loss, and damages. Although such evidence might be insufficient on its own to prove one or more of these elements, event study evidence that negates any of the first three elements implies that plaintiffs will be unable to establish entitlement to damages. These observations explain why event studies play such a central role in securities fraud litigation.

Loss causation and price impact have taken center stage at the pleading and class certification stages. If the failure to establish price impact is fatal to the plaintiffs’ case, the defendants benefit by making that challenge at the pleading stage, before the plaintiffs can obtain discovery,<sup>101</sup> or by preventing plaintiffs from obtaining the leverage of class certification.<sup>102</sup> Accordingly, much of the Supreme Court’s jurisprudence on loss causation and price impact has been decided in the context of pretrial motions.

*Basic* itself was decided on a motion for class certification. A key factor in the Court’s analysis was the critical role that a presumption of reliance would play in enabling the plaintiff to address Rule 23’s commonality requirement.<sup>103</sup> As the Court explained, “[r]equiring proof of individualized reliance from each member of the proposed plaintiff class effectively would

98. *Id.* at 672.

99. *Id.* at 671 (citing *Amgen*, 133 S. Ct. at 1203).

100. *Halliburton Co.*, 309 F.R.D. at 251. The parties subsequently agreed to a class settlement, and the district court issued an order preliminarily approving that settlement, pending a fairness hearing. *Erica P. John Fund, Inc. v. Halliburton Co.*, No. 3:02-CV-01152-M, at \*1 (N.D. Tex. Mar. 31, 2017).

101. Under the PSLRA, “all discovery and other proceedings shall be stayed during the pendency of any motion to dismiss” subject to narrow exceptions. 15 U.S.C. § 78u-4(b)(3)(B) (2010).

102. *See, e.g.*, Transcript of Oral Argument at 23, *Halliburton Co. v. Erica P. John Fund, Inc.* (*Halliburton II*), 134 S. Ct. 2398 (2014) (No. 13-317) (Justice Scalia: “Once you get the class certified, the case is over, right?”).

103. *Basic Inc. v. Levinson*, 485 U.S. 224, 242–43, 249 (1988).

have prevented respondents from proceeding with a class action, since individual issues then would have overwhelmed the common ones.”<sup>104</sup> By facilitating class certification, *Basic* has been described as transforming private securities fraud litigation.<sup>105</sup>

Defendants have responded by attempting to increase the burden imposed on the plaintiff to obtain class certification. In *Halliburton I*, the lower courts accepted defendant’s argument that plaintiffs should be required to establish loss causation at class certification.<sup>106</sup> In *Amgen*, the defendants argued that the plaintiff should be required to establish materiality in order to obtain class certification.<sup>107</sup> Notably, in both cases, the defendants’ objective was to require the plaintiffs to prove price impact through an event study at a preliminary stage in the litigation rather than at the merits stage.

Similarly, the Court’s decision in *Dura Pharmaceuticals* was issued in the context of a motion to dismiss for failure to state a claim.<sup>108</sup> The complaint ran afoul of even the pre-*Twombly*<sup>109</sup> pleading standard by failing to allege that there had been any corrective disclosure associated with a loss.<sup>110</sup> The *Dura* Court held that the plaintiffs’ failure to plead loss causation meant that the complaint did not show entitlement to relief as required under Rule 8(a)(2).<sup>111</sup> In the post-*Dura* state of affairs, plaintiffs must identify both alleged misrepresentation and corrective disclosure dates to adequately plead loss causation. They would also be well-advised to allege that an expert-run event study establishes materiality, reliance, loss causation, economic loss, and damages. Failure to do so would not necessarily be fatal, but it would leave plaintiffs vulnerable to a Rule 12(b)(6) motion to dismiss. Given the importance of the event study in securities litigation, it is important to understand both the methodology involved and its limitations.

## II. The Theory of Financial Economics and the Practice of Event Studies: An Overview

The theory of financial economics adopted by courts for purposes of securities litigation is based on the premise that publicly released information

104. *Id.* at 242.

105. *See, e.g.,* Langevoort, *supra* note 54, at 152 (“Tens of billions of dollars have changed hands in settlements of 10b-5 lawsuits in the last twenty years as a result of *Basic*.”).

106. *Archdiocese of Milwaukee Supporting Fund, Inc. v. Halliburton Co.*, 597 F.3d 330, 344 (5th Cir. 2010); *Archdiocese of Milwaukee Supporting Fund, Inc. v. Halliburton Co.*, No. 3:02-CV-1152-M, 2008 WL 4791492, at \*20 (N.D. Tex. Nov. 4, 2008).

107. *Amgen Inc. v. Conn. Ret. Plans & Trust Funds*, 568 U.S. 455, 459 (2013).

108. *Dura Pharm., Inc. v. Broudo*, 544 U.S. 336, 339–40 (2005).

109. *Bell Atl. Corp. v. Twombly*, 550 U.S. 544 (2007).

110. *Dura*, 544 U.S. at 347 (“[T]he complaint nowhere . . . provides the defendants with notice of what the relevant economic loss might be or of what the causal connection might be between that loss and the misrepresentation concerning *Dura*’s [product].”).

111. *Id.* at 346; FED. R. CIV. P. 8(a)(2).

concerning a security's price will be incorporated into its market price quickly.<sup>112</sup> This premise is known in financial economics as the semi-strong form of the "efficient market" hypothesis,<sup>113</sup> but we will refer to it simply as the efficient market hypothesis. Under the efficient market hypothesis, information that overstates a firm's value will quickly inflate the firm's stock price over the level that true conditions warrant. Conversely, information that corrects such inflationary misrepresentations will quickly lead the stock price to fall.

Financial economists began using event studies to measure how much stock prices respond to various types of news.<sup>114</sup> Typically, event studies focus not on the level of a stock's price, but on the percentage change in stock price, which is known as the stock's observed "return." In its simplest form, an event study compares a stock's return on a day when news of interest hits the market to the range of returns typically observed for that stock, taking account of what would have been expected given general changes in the overall market on that day. For example, if a stock typically moves up or down by no more than 1% in either direction but rises by 2% on a date of interest (after controlling for relevant market conditions), then the stock return moved an unusual amount on that date. What range is "typical," and thus how large must a return be to be considered sufficiently unusual, are questions that event study authors answer using statistical significance testing.

A typical event study has five basic steps: (1) identify one or more appropriate event dates, (2) calculate the security's return on each event date, (3) determine the security's expected return for each event date, (4) subtract the actual return from the expected return to compute the excess return for each event date, and (5) evaluate whether the resulting excess return is statistically significant at a chosen level of statistical significance.<sup>115</sup> We treat these five steps in two sections.

112. *Basic Inc. v. Levinson*, 485 U.S. 224, 245–47 (1988) ("[T]he market price of shares traded on well-developed markets reflects all publicly available information, and, hence, any material misrepresentations.").

113. There are also strong and weak forms. The strong form of the efficient market hypothesis holds that even information that is held only privately is reflected in stock prices since those with the information can be expected to trade on it. ROBERT L. HAGIN, *THE DOW JONES-IRWIN GUIDE TO MODERN PORTFOLIO THEORY* 12 (1979). The weak form holds only that "historical price data are efficiently digested and, therefore, are useless for predicting subsequent stock price changes." *Id.*

114. For a history of the use of event studies in academic scholarship, see A. Craig MacKinlay, *Event Studies in Economics and Finance*, 35 J. ECON. LITERATURE 13, 13–14 (1997).

115. Jonathan Klick & Robert H. Sitkoff, *Agency Costs, Charitable Trusts, and Corporate Control: Evidence from Hershey's Kiss-Off*, 108 COLUM. L. REV. 749, 798 (2008).



*A. Steps (1)–(4): Estimating a Security's Excess Return*

Experts typically address the first step (selecting the event date) by using the date on which the representation or disclosure was publicly made.<sup>116</sup> For purposes of public-market securities fraud, the information must be communicated widely enough that the market price can be expected to react to the information.<sup>117</sup> The second step (calculating a security's actual return) requires only public information about daily security prices.<sup>118</sup>

The third step is to determine the security's expected return on the event date, given market conditions that might be expected to affect the firm's price even in the absence of the news at issue. Event study authors do this by using statistical methods to separate out components of a security's return that are based on overall market conditions from the component due to firm-specific information. Market conditions typically are measured using a broad index of other stocks' returns on each date considered in the event study or an index of returns of other firms engaged in similar business (since firms engaged in common business activities are likely to be affected by similar types of information). To determine the expected return for the security in question, an expert will estimate a regression model that controls for the returns to market or industry stock indexes.<sup>119</sup> The estimated coefficients from this model can then be used to measure the expected return for the firm in question, given the performance of the index variables included in the model.

---

116. The event study literature contains an extensive treatment of the appropriate choice of event window, a topic that we do not consider in detail here. See Allen Ferrell & Atanu Saha, *The Loss Causation Requirement for Rule 10b-5 Causes of Action: The Implications of Dura Pharmaceuticals, Inc. v. Broudo*, 63 BUS. LAW. 163, 167–68 (2007) (discussing factors affecting choice of event window); Rinaudo & Saha, *supra* note 17, at 163 (observing that the typical event window is a single day but advocating instead for an “intraday event study methodology relying on minute-by-minute stock price data”). The choice of window may play a critical role in determining the results of the event study. See, e.g., *In re Intuitive Surgical Sec. Litig.*, No. 5:13-cv-01920-EJD, 2016 WL 7425926, at \*14 (N.D. Cal. Dec. 22, 2016) (holding the defendants' expert's usage of a two-day window was inappropriate and going on to find that the defendants failed to rebut plaintiffs' presumption of reliance).

117. In some cases, litigants may dispute whether information is sufficiently public to generate a market reaction; in other situations, leakage of information before public announcement may generate an earlier market reaction. See *Sherman v. Bear Stearns Cos.* (*In re Bear Stearns Cos., Sec., Derivative, & ERISA Litig.*), No. 09 Civ. 8161 (RWS), 2016 U.S. Dist. LEXIS 97784, at \*20–23 (S.D.N.Y. 2016) (describing various decisions analyzing the “leakage analysis”). These specialized situations can be addressed by tailoring the choice of event date.

118. Recall that a security's daily return on a particular date is the percentage change in the security over the preceding date.

119. As one pair of commentators has recently noted: “The failure to make adjustments for the effect of market and industry moves nearly always dooms an analysis of securities prices in litigation.” Brav & Heaton, *supra* note 11, at 590.

The fourth step is to calculate the “excess return,”<sup>120</sup> which one does by subtracting the expected return from the actual return on the date in question. Thus the excess return is the component of the actual return that cannot be explained by market movements on the event date, given the regression estimates described above. So the excess return measures the stock’s reaction to whatever news occurred on the event date.

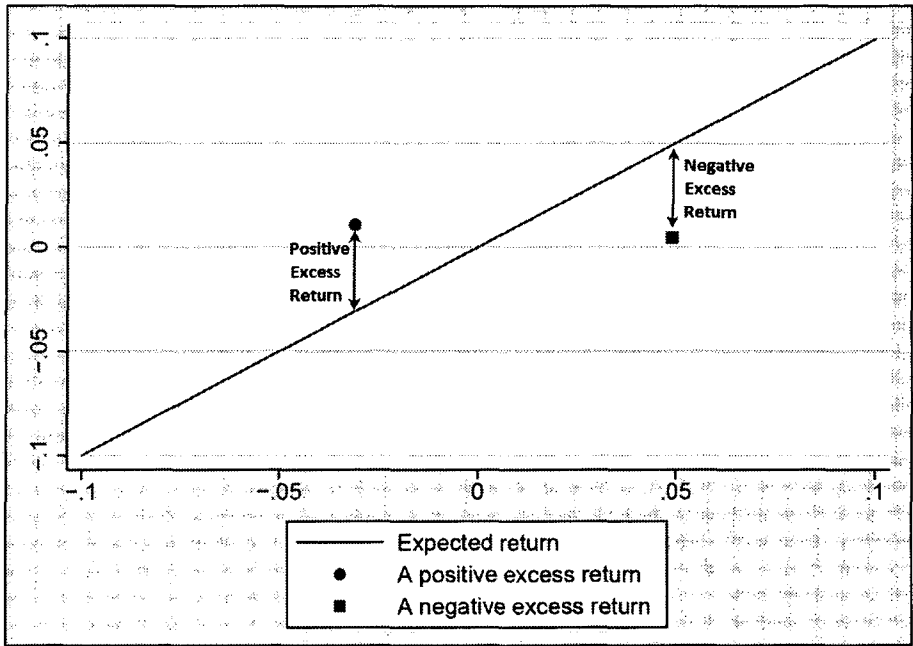
A positive excess return indicates that the firm’s stock increased more than would be expected based on the statistical model. A negative excess return indicates that the stock fell more than the model predicts it should have. Figure 1 illustrates the calculation of excess returns from actual returns and expected returns. The figure plots the stock’s actual daily return on the vertical axis and its expected daily return on the horizontal axis. The upwardly sloped straight line represents the collection of points where the actual and expected returns are equal. The magnitude of the excess return at a given point is the height between that point and the upwardly sloped straight line. The point plotted with a circle lies above the line where actual and expected returns are equal, so this point indicates a positive excess return. By contrast, at the point plotted with a square, the actual return is below the line where the actual and expected returns are equal, so the excess return is negative.

---

120. The term “abnormal return” is interchangeable with excess return. We use only “excess return” in this Article in order to avoid confusing “abnormal returns” with non-normality in the distribution of these returns.



Figure 1: Illustrating the Calculation of Excess Returns from Actual and Expected Returns



B. Step (5): Statistical Significance Testing in an Event Study

Our fifth and final step is to determine whether the estimated excess return is statistically significant at the chosen level of significance, which is frequently the 5% level. The use of statistical significance testing is designed to distinguish stock-price changes that are just the result of typical volatility from those that are sufficiently unusual that they are likely a response to the alleged corrective disclosure.

Tests of statistical significance all boil down to asking whether some statistic’s observed value is far enough away from some baseline level one would expect that statistic to take. For example, if one flips a fair coin 100 times, one should expect to see heads come up on roughly 50% of the flips, so the baseline level of the heads share is 50%. The hypothesis that the coin is fair, so that the chance of a heads is 50%, is an example of what statisticians call a *null hypothesis*: a maintained assumption about the object of statistical study that will be dropped only if the statistical evidence is sufficiently inconsistent with the assumption.

Since one can expect random variation to affect the share of heads in 100 coin flips, most scholars would find it unreasonable to reject the null hypothesis that the coin is fair simply because one observes a heads share of,

say, 49% or 51%. Even though these results do not equal exactly the baseline level, they are close enough that most applied statisticians would consider this evidence too weak to reject the null hypothesis that the coin is fair.<sup>121</sup> On the other hand, common sense and statistical methodology suggest that if eighty-nine of 100 tosses yielded heads, it would be strong evidence that the coin was biased toward heads. A finding of eighty-nine heads would cause most scholars to reject the null hypothesis that the coin is fair.

Event study tests of whether a stock price moved in response to information are similar to the coin toss example. They seek to determine whether the stock's excess return was highly unusual on the event date. The null hypothesis in an event study is that the news at issue did not have any price impact. Under this null hypothesis, the stock's return should reflect only the usual relationship between the stock and market conditions on the event date. In other words, the stock's return should be the expected return, together with normal variation. Our baseline expectation for the stock's excess return is that it should be zero. Normal variation, however, will cause the stock's actual return to differ somewhat from the expected return. Statistical significance testing focuses on whether this deviation—the actual excess return on the event date—is highly unusual.

What counts as highly unusual in securities litigation? Typically courts and experts have treated an event-date effect as statistically significant if the event-date's excess return is among the 5% most extreme values one would expect to observe in the absence of any fraudulent activity.<sup>122</sup> In this situation,

---

121. At the same time, observing a heads share of 49% does provide some weak evidence that the coin is biased toward tails. A simple way to quantify that evidence is to use a result based on Bayes' theorem, according to which the posterior odds in favor of a proposition equal the product of the prior odds and the likelihood ratio. *See, e.g.,* David H. Kaye & George Sensabaugh, *Reference Guide on DNA Identification Evidence*, in REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 129, 173 (3d ed. 2011) (describing Bayes' theorem). Whatever the prior odds that the coin in favor of a true heads probability equal to 0.49, the likelihood ratio in favor of this proposition will exceed 1 since the observed data are more likely when the heads probability is 0.49 than when it is 0.5. When the likelihood ratio exceeds 1, the posterior odds exceed the prior odds, so the data provide some support for the alternative hypothesis of a coin that is slightly biased toward tails. A more complete discussion of this issue would have to address the question of the prior probability distribution over non-fair heads probabilities, which involves replacing the numerator of the likelihood ratio with its average over the prior distribution (the resulting ratio is known as the Bayes factor). The dominant approach to applied statistics among scholars, and certainly among experts in litigation, is the frequentist approach, which is usually hostile to the specification of priors. That is why frequentists focus on statistical significance testing rather than reporting posterior odds or probabilities. Further details are beyond the scope of the present Article.

122. *See, e.g.,* Erica P. John Fund, Inc. v. Halliburton Co., 309 F.R.D. 251, 262 (N.D. Tex. 2015) ("To show that a corrective disclosure had a negative impact on a company's share price, courts generally require a party's expert to testify based on an event study that meets the 95% confidence standard . . ." This standard requires that "one can reject with 95% confidence the null hypothesis that the corrective disclosure had no impact on price.") (citing Fox, *supra* note 92, at 442 n.17); *cf.* Brav & Heaton, *supra* note 11, at 596–99 (questioning whether requiring statistical significance at the 95% confidence level for securities fraud event studies is appropriate). The genesis of the 5% significance level is most probably its use by R.A. Fisher in his influential

experts equivalently say that there is statistically significant evidence at the 5% level, or “at level 0.05,” or “with 95% confidence.”<sup>123</sup>

Implicit in this discussion of statistical significance is the scholarly norm of declaring that evidence that disfavors a null hypothesis is not strong enough to reject that hypothesis. Thus, applied statisticians often say that a statistically insignificant estimate is not necessarily proof that the null hypothesis is true—just that the evidence isn’t strong enough to declare it false. Such statisticians really have three categories of conclusion: that the evidence is strong enough to reject the null hypothesis, that the evidence is basically consistent with the null hypothesis, and that the evidence is inconsistent with the null hypothesis but not so much as to warrant rejection of the null hypothesis. One might think of such statisticians who use demanding significance levels such as the 5% level as starting with a strong presumption in favor of the null hypothesis so that only strong evidence against it will be deemed sufficient to reject the null hypothesis.

Whether an approach of adopting a strong presumption in favor of the defendant is consistent with legal standards in securities litigation is beyond the scope of this Article but it is a topic that warrants future discussion.<sup>124</sup> For purposes of this Article, though, we take the choice of the 5% significance level as given and seek to provide courts with the methodological knowledge necessary to apply that significance level properly.<sup>125</sup>

Experts typically assume that in the absence of any fraud-related event, a stock’s excess returns—that is, the typical variability not driven by the news at issue in litigation—will follow a normal distribution,<sup>126</sup> an issue we discuss in more detail in Part IV. For a random variable that follows a normal distribution, 95% of realizations of that variable will take on a value that is

---

textbook. See R.A. FISHER, *STATISTICAL METHODS FOR RESEARCH WORKERS* 45, 85 (F.A.E. Crew & D. Ward Cutler eds., 5th ed. 1934).

123. That is not to say that the event study can determine whether this price effect is rational in the substantive sense that Justice Alito seems to have had in mind. See Transcript of Oral Argument at 24, *Halliburton Co. v. Erica P. John Fund, Inc. (Halliburton II)*, 134 S. Ct. 2398 (2014) (No. 13-317) (asking whether event studies can determine market irrationality). The measured price impact represented by the excess return is simply the effect that is empirically evident from investor behavior in the relevant financial market.

124. For a discussion of some of these issues outside the securities litigation context, see Michelle M. Burtis, Jonah B. Gelbach & Bruce H. Kobayashi, *Error Costs, Legal Standards of Proof and Statistical Significance* 2–7, 9–14 (George Mason Law & Econ. Research Paper No. 17-21, 2017), <https://ssrn.com/abstract=2956471> [<https://perma.cc/FRJ3-FNX7>].

125. *Daubert* requires at least this much. *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579, 590–91 n.9 (1993) (equating evidentiary reliability of scientific testimony with scientific validity and defining scientific validity as the requirement that a “principle support[s] what it purports to show”).

126. See, e.g., Brav & Heaton, *supra* note 11, at 591 n.17 (“[S]tandard practice still rests heavily on the normality assumption . . .”).

within 1.96 standard deviations of zero.<sup>127</sup> Experts assuming normality of excess returns and using the 95% confidence level often determine that the excess return is highly unusual if it is greater than 1.96 standard deviations. For example, if the standard deviation of a stock's excess returns is 1.5%, an expert might declare an event date's excess return statistically significant only if it is more than 2.94 percentage points from zero.<sup>128</sup> In this example, the expert has determined that the "critical value" is 2.94: any value of the event date excess return greater in magnitude than this value will lead the expert to determine that the excess return is statistically significant at the 5% level. A lower value for the excess return would lead to a finding of statistical insignificance.

When an event date excess return is statistically significant at the chosen significance level, courts will treat the size of the excess return as a measure of the price effect associated with the news at issue.<sup>129</sup> One consequence is that the excess return may then be used as a basis for determining damages. On the other hand, if the excess return is statistically insignificant at the chosen level, then courts find the statistical evidence too weak to meet the plaintiff's burden of persuasion that the information affected the stock price.

Note that a statistically insignificant finding may occur even when the excess return is directionally consistent with the plaintiff's allegations. In such a case, the evidence is consistent with the plaintiff's theory of the case, but the size of the effect is too small to be statistically significant at the level used by the court. Such an outcome may sometimes occur even when the null hypothesis was really false, i.e., there really was a price impact due to the news on the event date.

This last point hints at an inherent trade-off reflected in statistical significance testing. When one conducts a statistical significance test, there are four possible outcomes. These four categories of statistical inference are summarized in Table 1. Two of these are correct inferences: the test may fail to reject a null hypothesis that is really true, or the test may reject a null hypothesis that is really false. The first of these cases correctly determines that there was no price impact (the upper left box in Table 1). The second case correctly determines that there was a price impact (the lower right box

---

127. The standard deviation is a measure of how spread out a large random sample of the variable is likely to be. The standard deviation of a firm's excess returns is often estimated using the root-mean-squared error, a statistic that is usually reported by statistical software. *See, e.g.*, HUMBERTO BARRETO & FRANK M. HOWLAND, *INTRODUCTORY ECONOMETRICS: USING MONTE CARLO SIMULATION WITH MICROSOFT EXCEL 117* (2006) (describing the calculation and use of root-mean-squared error).

128. This figure arises because 1.96 times 1.5 is 2.94. As we discuss in Part IV, *infra*, there are a number of potential problems with this typical approach.

129. *See Brav & Heaton, supra* note 11, at 600–01 (explaining that many courts applying the event study approach look to the size of the excess return in relation to a predetermined statistical significance level to determine whether the price impact is actionable).

in Table 1). Given that there really was a price impact, the probability of correctly making this determination is known as the test’s power.<sup>130</sup>

The other two outcomes are incorrect inferences. The first mistaken inference involves rejecting a null hypothesis that is actually true. This is known as a Type I error (top right box in Table 1). The probability of this result, given that the null hypothesis is true, is known as a test’s size.<sup>131</sup> The second incorrect inference is failing to reject a null hypothesis that is actually false (lower left box in Table 1); this is known as a Type II error.<sup>132</sup>

Table 1: Four Categories of Statistical Inference

	<u>Don’t Reject Null</u> Test does not find statistically significant price effect	<u>Reject Null</u> Test finds statistically significant price effect
<u>Null is true</u> No highly unusual price effect	Accurate finding of no price effect	Type I error (Size)
<u>Null is false</u> Highly unusual price effect	Type II error	Accurate finding of price effect (Power)

The trade-off that arises in statistical significance testing is simple: reducing a test’s Type I error rate means increasing its Type II error rate, and vice versa.<sup>133</sup> As noted above, event study authors usually use a confidence

130. Thus power is the probability of winding up in the lower right box in Table 1, given that we must wind up in one of the two lower boxes; it is the ability of the test to identify a price impact when it actually exists.

131. For this reason, a test with significance level of 5% is sometimes said to have size 0.05.

132. Given that the null hypothesis is false so that we must wind up in one of the two lower boxes in Table 1, the probability of a Type II error equals one minus the test’s power. *See Brav & Heaton, supra* note 11, at 593 & n.26 (“Statistical power describes the probability that a test will correctly identify a genuine effect.” (quoting PAUL D. ELLIS, *THE ESSENTIAL GUIDE TO EFFECT SIZES: STATISTICAL POWER, META-ANALYSIS, AND THE INTERPRETATION OF RESEARCH RESULTS* 52 (2010))).

133. To be sure, it is sometimes true that two tests have the same Type I error rate but different Type II error rates (or vice versa). However, the Type II error rate for a given test—such as the significance testing approach typically used in event studies—can be reduced only by increasing the Type I error rate (and vice versa).

level of 95%, which is the same as a Type I error rate of 5%.<sup>134</sup> The Type II error rate associated with this Type I error rate will depend on the typical range of variability of excess returns, but it has recently been pointed out that insisting on a Type I error rate of 5% when using event studies in securities fraud litigation can be expected to cause very high Type II error rates.<sup>135</sup> Another way to put this is that event studies used in securities litigation are likely to have very low power—very low probability of rejecting an actually false null hypothesis—when we insist on keeping the Type I error rate as low as 5%.<sup>136</sup> We discuss this very important issue further in subpart V(C).

A final issue related to statistical significance concerns who bears the burden of persuasion if the defendant seeks to use event study evidence to show that there was no price impact related to an alleged misrepresentation. *Halliburton II* states that “defendants must be afforded an opportunity before class certification to defeat the presumption through evidence that an alleged misrepresentation did not actually affect the market price of the stock.”<sup>137</sup> But the case does not announce what statistical standard will apply to defendants’ evidence. As Merritt Fox discusses, one view is that the defendant must present statistically significant evidence that the price changed in the direction *opposite* to the plaintiff’s allegations.<sup>138</sup> Alternatively, the defendant might have to present evidence that is sufficient only to persuade the court that its own evidence of the absence of price impact is more persuasive than the plaintiff’s affirmative evidence of price impact.<sup>139</sup>

As Fox has noted in other work, the applicable legal standard will have considerable impact on the volume of cases that are able to survive beyond a preliminary stage.<sup>140</sup> Further, Fox points out, a variety of factors affect the choice of approach, including social policy considerations about the appropriate volume of securities fraud litigation.<sup>141</sup> The question of Rule 301’s applicability was appealed to the Fifth Circuit by the *Halliburton* parties, but the parties reached a proposed settlement before that court could issue its ruling.<sup>142</sup> A full discussion of these issues is beyond the scope of the

134. See, e.g., *In re Intuitive Surgical Sec. Litig.*, No. 5:13-cv-01920-EJD, 2016 WL 7425926, at \*15 (N.D. Cal. Dec. 22, 2016).

135. See Brav & Heaton, *supra* note 11, at 593–97 (demonstrating that, as a result, the standard event study will frequently fail to reject the null hypothesis when the actual price impact is small).

136. For an excellent in-depth discussion, see *id.*

137. *Halliburton Co. v. Erica P. John Fund, Inc. (Halliburton II)*, 134 S. Ct. 2398, 2417 (2014).

138. See Fox, *supra* note 92, at 447–49.

139. *Id.* at 454–55. As Fox discusses, Federal Rule of Evidence 301 provides some support for this second approach. *Id.* at 457. However, Fox also points out a number of complicating issues as to the applicability of Rule 301 to 10b-5 actions. *Id.* at 457–58.

140. Merritt B. Fox, *Halliburton II: What It’s All About*, 1 J. FIN. REG. 135, 139–41 (2015).

141. *Id.* at 141.

142. See, e.g., Brief of Appellants *Halliburton Co. & David J. Lesar* at 52–60, *Erica P. John Fund, Inc. v. Halliburton Co.*, No. 15-11096 (5th Cir. filed Feb. 8, 2016) (arguing that FED. R. EVID. 301 applies and “dictate[s] that plaintiffs bear the burden of persuasion on price impact”); Brief of



present Article. For concreteness, we will simply follow the approach taken by the district court in the ongoing *Halliburton* litigation. While that court found “that both the burden of production and the burden of persuasion are properly placed on Halliburton,”<sup>143</sup> the court did not understand that burden allocation to require Halliburton to affirmatively disprove the plaintiff’s allegations statistically. Rather, Halliburton needed only to “persuade the Court that its expert’s event studies [were] more probative of price impact than the Fund’s expert’s event studies.”<sup>144</sup> The rest of the court’s opinion makes clear that this means treating both sides’ event studies as if they are testing whether the statistical evidence is sufficient to establish that there is statistically significant evidence of a price impact at the 5% level, as discussed above. We will therefore continue to concentrate on that approach throughout this Article.

The foregoing discussion summarizes the basic methodology of event studies as they are commonly used in securities litigation. In the next Part, we present our own stylized event study of dates involved in the ongoing *Halliburton* litigation both to illustrate the principles described above and to facilitate our Part IV discussion of important refinements that experts and courts should make to achieve consistency with announced standards. We raise the question of whether those standards are appropriate in Part V.

### III. The Event Study as Applied to the *Halliburton* Litigation

This Part uses data and methods from the opinions and expert reports in the *Halliburton* case to illustrate and critically analyze the use of an event study to measure price impact. Our objective is, initially, to provide a basic application of the theory described in the preceding Part for those readers having limited familiarity with the operational details. Then, in Part IV, we identify several problems with the typical execution of the basic approach and demonstrate the implications of making the necessary adjustments to respond to these problems.

#### A. *Dates and Events at Issue in the Halliburton Litigation*

Plaintiffs in the *Halliburton* litigation alleged that between the middle of 1999 and the latter part of 2001,<sup>145</sup> Halliburton and several of the

---

the Lead Plaintiff-Appellee & the Certified Class at 49–58, *Erica P. John Fund, Inc. v. Halliburton Co.*, No. 15-11096 (5th Cir. filed Mar. 28, 2016) (contending that Rule 301 does not apply to relieve Halliburton of its burden of production and persuasion); as to settlement, see *Erica P. John Fund, Inc. v. Halliburton Co.*, No. 3:02-CV-01152-M, at \*1 (N.D. Tex. Mar. 31, 2017).

143. *Erica P. John Fund, Inc. v. Halliburton Co.*, 309 F.R.D. 251, 260 (N.D. Tex. 2015).

144. *Id.*

145. We focus on the class period at issue at the time of the most recent district court order, which ran from July 22, 1999, to December 7, 2001. The class period referred to in the operative complaint began slightly earlier, on June 3, 1999. Fourth Consolidated Amended Complaint for Violation of the Securities Exchange Act of 1934 para. 1, Archdiocese of Milwaukee Supporting

company's officers—collectively referred to here as simply “Halliburton”—made false and misleading statements about various aspects of the company's business.<sup>146</sup> The operative complaint, together with the report filed by plaintiffs' experts, named a total of thirty-five dates on which either misrepresenting statements or corrective disclosures (or both) allegedly occurred.<sup>147</sup> For purposes of illustration, consider two of the allegedly fraudulent statements:

- (1) Plaintiffs alleged that in a 1998 10-K report filed on March 23, 1999, Halliburton failed to disclose that it faced the risk of having to “shoulder the responsibility” for certain asbestos claims filed against other companies; further, plaintiffs alleged that Halliburton failed to correctly account for this risk.<sup>148</sup>
- (2) On November 8, 2001, Halliburton stated in its Form 10-Q filing for the third quarter of 2001 that the company had an accrued liability of \$125 million related to asbestos claims and that “[W]e believe that open asbestos claims will be resolved without a material adverse effect on our financial position or the results of operations.”<sup>149</sup> Plaintiffs also alleged that this representation was false and misleading.<sup>150</sup>

Both the alleged misrepresentations described above were confirmatory in the sense that the plaintiffs alleged that Halliburton, rather than accurately informing the market of negative news, falsely confirmed prior good news that was no longer accurate.<sup>151</sup> The alleged result was that Halliburton's stock price was inflated because it remained at a higher level than it would have had Halliburton disclosed accurately. Since false confirmatory misrepresentations do not constitute “new” information—even under the plaintiffs' theory—neither of the two statements above would have been expected to cause an increase in Halliburton's market price. As a result, in considering the price impact of the alleged misrepresentations, the district

---

Fund, Inc. v. Halliburton Co., No. 3:02-CV-1152-M (N.D. Tex. filed Apr. 4, 2006) [hereinafter FCAC]. The difference is immaterial for our purposes.

146. *Id.* ¶ 2.

147. *Halliburton Co.*, 309 F.R.D. at 264. A defense expert report lists twenty-five distinct dates on which plaintiffs or their expert alleged misrepresentations. Expert Report of Lucy P. Allen ¶ 10, Archdiocese of Milwaukee Supporting Fund, Inc. v. Halliburton Co., No. 3:02-CV-1152-M (N.D. Tex. filed Sept. 10, 2014) [hereinafter Allen Report].

148. FCAC, *supra* note 145, ¶ 74.

149. *Id.* ¶ 189.

150. *Id.* ¶ 190.

151. See Archdiocese of Milwaukee Supporting Fund, Inc., v. Halliburton Co., No. 3:02-CV-1152-M, 2008 U.S. Dist. LEXIS 89598, at \*17–18 (N.D. Tex. 2008) (discussing the “[p]laintiffs['] claim that Halliburton made material misrepresentations . . . to inflate the price of [its] stock”).



court allowed the plaintiffs to focus on whether subsequent alleged corrective disclosures were associated with reductions in Halliburton's stock price.<sup>152</sup>

On July 25, 2015, the district court issued its most recent order and memorandum opinion concerning class certification.<sup>153</sup> By this point of the litigation, which had been ongoing for more than thirteen years, the event studies submitted by the parties' experts<sup>154</sup> focused on six dates on which Halliburton had issued alleged corrective disclosures: December 21, 2000;<sup>155</sup> June 28, 2001;<sup>156</sup> August 9, 2001;<sup>157</sup> October 30, 2001;<sup>158</sup> December 4, 2001;<sup>159</sup> and December 7, 2001.<sup>160</sup>

The trial court concluded in its July 2015 decision, after weighing two competing expert reports, that five of these alleged corrective disclosures did

152. *Halliburton Co.*, 309 F.R.D. at 262 ("Measuring price change at the time of the corrective disclosure, rather than at the time of the corresponding misrepresentation, allows for the fact that many alleged misrepresentations conceal a truth."). As discussed in Part I, this is not a novel approach. For example, one court of appeals has explained:

[P]ublic statements falsely stating information which is important to the value of a company's stock traded on an efficient market may affect the price of the stock even though the stock's market price does not soon thereafter change. For example, if the market believes the company will earn \$1.00 per share and this belief is reflected in the share price, then the share price may well not change when the company reports that it has indeed earned \$1.00 a share even though the report is false in that the company has actually lost money (presumably when that loss is disclosed the share price will fall).

*Nathenson v. Zonagen Inc.*, 267 F.3d 400, 419 (5th Cir. 2001). In contrast, by its very nature a corrective disclosure cannot be confirmatory: for the alleged corrective disclosure to be truly corrective, it must really be new news. Thus, evidence concerning the stock price change on the date of an alleged corrective disclosure will always be probative. For simplicity, we will generally focus on the case in which alleged misrepresentations were confirmatory, leading us to analyze the corrective disclosure date. *But see* section IV(C)(3), *infra*, which considers the situation when plaintiffs must establish price impact on both an alleged misrepresentation date and an alleged corrective disclosure date.

153. *Halliburton Co.*, 309 F.R.D. at 280.

154. Expert Report of Chad Coffman, CFA, Archdiocese of Milwaukee Supporting Fund, Inc. v. Halliburton Co., No. 3:02-CV-1152-M, 2008 U.S. Dist. LEXIS 89598 (N.D. Tex. 2008) [hereinafter Coffman Report] (plaintiffs' expert); Allen Report, *supra* note 147 (defendants' expert).

155. On this date, "Halliburton announced a \$120 million charge which included \$95 million in project costs, some of which allegedly should not have been previously booked." Coffman Report, *supra* note 154, ¶ 8 (citing FCAC, *supra* note 145, ¶ 150).

156. On this date, "Halliburton disclosed that" third-party "Harbison-Walker asked for asbestos claims related financial assistance from Halliburton." *Id.* (citing FCAC, *supra* note 145, ¶ 170).

157. On this date, Halliburton's "2Q01 10-Q included additional details regarding asbestos claims." *Id.* (citing FCAC, *supra* note 145, ¶ 178).

158. On this date, "Halliburton issued a press release announcing the Mississippi verdict." *Id.* (citing *Form 8-K*, HALLIBURTON (Nov. 6, 2001), [http://ir.halliburton.com/phoenix.zhtml?c=67605&p=irol-sec&seccat01enhanced.1\\_rs=11&seccat01enhanced.1\\_rc=10](http://ir.halliburton.com/phoenix.zhtml?c=67605&p=irol-sec&seccat01enhanced.1_rs=11&seccat01enhanced.1_rc=10) [<https://perma.cc/A9U4-8QSK>]).

159. On this date, "Halliburton announced Texas judgment and three other judgments." *Id.* (citing FCAC, *supra* note 145, ¶ 191).

160. On this date, "Halliburton announced Maryland verdict." *Id.* (citing FCAC, *supra* note 145, ¶ 191).

not have a price impact that was statistically significant at the 5% level. For that reason, the district court denied class certification with respect to these five dates.<sup>161</sup> However, the district court found that the alleged corrective disclosure on December 7 was associated with a statistically significant price impact at the 5% level, in the direction necessary for plaintiffs to benefit from the *Basic* presumption. The court therefore certified a class action with respect to the alleged misrepresentations associated with December 7, 2001.<sup>162</sup>

*B. An Illustrative Event Study of the Six Dates at Issue in the Halliburton Litigation*

Following the approach outlined in Part II, we apply the event study to the six dates listed in subpart III(A). For our first step (selection of an appropriate event), we follow the parties and analyze the dates of the alleged corrective disclosures.<sup>163</sup>

Next, we use the market model to construct Halliburton's estimated return.<sup>164</sup> To account for factors outside the litigation likely associated with Halliburton's stock performance, we followed the parties' experts and estimated a market model with multiple reference indexes. The first such index, introduced by the defendants' expert, is intended to track the performance of the S&P 500 Energy Index during the class period.<sup>165</sup> The

161. Erica P. John Fund, Inc. v. Halliburton Co., 309 F.R.D. 251, 279–80 (N.D. Tex. 2015).

162. *Id.* at 280. Halliburton subsequently requested and received permission to pursue an interlocutory appeal of the class certification order pursuant to Rule 23(f). Erica P. John Fund, Inc. v. Halliburton Co., No. 15–90038, 2015 U.S. App. LEXIS 19519, at \*3 (5th Cir. Nov. 4, 2015). The issues on appeal did not concern the statistical aspects of event study evidence but rather were related to the district court's determination that Halliburton could not, at the class certification stage, provide nonstatistical evidence challenging the status of news as a corrective disclosure. *See id.* at \*1–2 (Dennis, J. concurring) (“The petition raises the question of whether a defendant in a federal securities fraud class action may rebut the presumption of reliance at the class certification stage by producing evidence that a disclosure preceding a stock-price decline did not correct any alleged misrepresentation.”). A settlement is pending in the case. Erica P. John Fund, Inc. v. Halliburton Co., No. 3:02-CV-01152-M, at \*1 (N.D. Tex. Mar. 31, 2017).

163. We do not independently address the legal question as to whether the disclosures made on the designated event dates are appropriately classified as corrective disclosures, as the trial court determined that whether a disclosure was correctly classified as corrective was not properly before the court at the class certification stage. *See Halliburton*, 309 F.R.D. at 261–62 (“[T]he issue of whether disclosures are corrective is not a proper inquiry at the certification stage.”).

164. Since the possibility of unusual stock return behavior is the object of an event study in the case, these dates should be removed from the set used in estimating the market model, and we do exclude them. This issue was controverted between the parties, with the plaintiffs' expert, Coffman, excluding all thirty-five of the dates identified in either the complaint or in an earlier expert's report. The district court accepted the argument that dates not identified as alleged corrective disclosure dates should be included in the event study, as defendants' expert had argued. *Id.* at 265.

165. The defendants' expert used this index in the market model, which she described in several reports. Allen Report, *supra* note 147, ¶ 20. We obtained a list of companies represented in this index during the class period from Exhibit 1 of the report of the plaintiffs' expert. Coffman Report,

plaintiffs' expert pointed out that this index is dominated by "petroleum refining companies, not energy services companies like Halliburton."<sup>166</sup> In his own market model, he therefore added a second index intended to reflect the performance of Halliburton's industry peers.<sup>167</sup> We also included such an index.<sup>168</sup> Third, we included an index constructed to mimic the one the defendants' expert constructed to reflect the engineering and construction aspects of Halliburton's business.<sup>169</sup> Because we found that the return on the S&P 500 overall index added no meaningful explanatory power to the model, we did not include it.

The resulting market model estimates<sup>170</sup> are set forth in Table 2.<sup>171</sup> These estimates indicate that Halliburton's daily stock return moves nearly one-for-one with the industry peer index constructed from analyst reports—a one percentage point increase in the industry peer index return is associated with roughly a 0.9-point increase in Halliburton's return. This makes the industry peer index a good tool for estimating Halliburton's expected return in the absence of fraud. The energy index return is much less correlated with Halliburton's stock return, with a coefficient of only about 0.2. Both the energy and industry peer index coefficients are highly statistically significant, with each being many multiples of its estimated standard error. By contrast, the return on the energy and construction index has essentially no association with Halliburton's stock return and is statistically insignificant.

---

*supra* note 154, at Exhibit 1. We then calculated the return on a value-weighted index based on these firms by calculating the daily percentage change in total market capitalization of these firms.

166. Coffman Report, *supra* note 154, ¶ 28.

167. This index is composed "of the companies cited by analysts as Halliburton's peers at least three times during the Class Period and with a market cap of at least \$1 billion at the end of the Class Period." *Id.* ¶ 33.

168. We calculated the return on this index in the same way as the return on the energy index described in note 165, *supra*; we took the list of included companies from Exhibit 3b of the Coffman Report. *Id.* at Exhibit 3b.

169. We took the list of companies for this index from the Allen Report, *supra* note 147, ¶ 20 n.20.

170. These estimates are calculated using the ordinary least squares estimator.

171. We used simple daily returns to estimate this model. We found nearly identical results when we entered all return variables in this model in terms of the natural logarithm of one plus the daily return, as experts sometimes do. For simplicity we decided to stick with the raw daily return.

Table 2: Market Model Regression Estimates

Variable	Coefficient Estimate	Estimated Standard Error
Industry Peer Index	0.903	0.031
Energy Index	0.210	0.048
E&C Index	0.033	0.036
Intercept	-0.001	0.001
Root mean squared error	1.745%	
Number of dates	593	

We then use these market model coefficient estimates to calculate daily estimated excess returns for the six event dates excluded from estimation of the model. We calculated the contribution of each index to each date’s expected return by multiplying the index’s Table 2 coefficient estimate by the observed value of the index on the date in question. Then we summed up the three index-specific products just created and added the intercept (which is so low as to be effectively zero). The result is the event date expected return based on the market model, i.e., the variable plotted on the horizontal axis of Figure 1 and Figure 3. The excess return for each event date is then found by subtracting each date’s estimated expected return from its actual return. Table 3 reports the actual, estimated expected, and estimated excess returns for each of the six alleged corrective disclosure dates in the *Halliburton* litigation, sorted from most negative to least negative. The actual returns are all negative, indicating that Halliburton’s stock price dropped on each of the alleged corrective disclosure dates. On three of the dates, the estimated expected return was also negative, indicating that typical market factors would be expected to cause Halliburton’s stock price to fall, even in the absence of any unusual event. For the other three dates, market developments would have been expected to cause an increase in Halliburton’s stock price. This means the estimated excess returns on those dates will imply larger price drops than are reflected in the actual returns. Finally, the estimated excess return column in Table 3 shows that the estimated excess returns were negative on all six dates. Even on dates when Halliburton’s stock price would have been expected to fall based on market developments, it fell *more* than it would have been expected to.

Table 3: Actual, Expected, and Excess Returns for Event Dates

Event Date	Actual Return	Estimated Expected Return	Estimated Excess Return
December 7, 2001	-42.4%	0.3%	-42.7%
August 9, 2001	-4.5%	0.6%	-5.1%
December 4, 2001	-0.7%	2.9%	-3.6%
December 21, 2000	-2.0%	-0.8%	-1.2%
October 30, 2001	-5.2%	-4.3%	-0.9%
June 28, 2001	-3.8%	-3.1%	-0.8%

The next step is to test these estimated excess returns for statistical significance in order to determine whether they are unusual enough to meet the court's standard for statistical significance.

For the moment, we adopt the standard assumption that Halliburton stock's excess returns follow a normal distribution. Our Table 2 above reports that the root-mean-squared error for our Halliburton market model—which is an estimate of the standard deviation of excess returns—was 1.745%. Multiplying 1.96 and 1.745, we obtain a critical value of 3.42%.<sup>172</sup> In other words, in the absence of unusual events affecting Halliburton's stock price and assuming normality, we can expect that 95% of Halliburton's excess returns will take on values between -3.42% and 3.42%. For an alleged corrective disclosure date, excess returns must be negative to support the plaintiff's theory, so a typical expert would determine that an event-date excess return drop of 3.42% or more is statistically significant.

In the first column of Table 4, we again present the estimated excess returns from Table 3. The second column reports whether the estimated excess return is statistically significant at the 5% level based on the standard approach to testing described above. The event date estimated excess returns are statistically significant at the 5% level for December 7, 2001; August 9, 2001; and December 4, 2001; they are statistically insignificant at the 5% level for the other three dates.

---

172. This follows because 1.96 times 1.745 equals 3.4202.

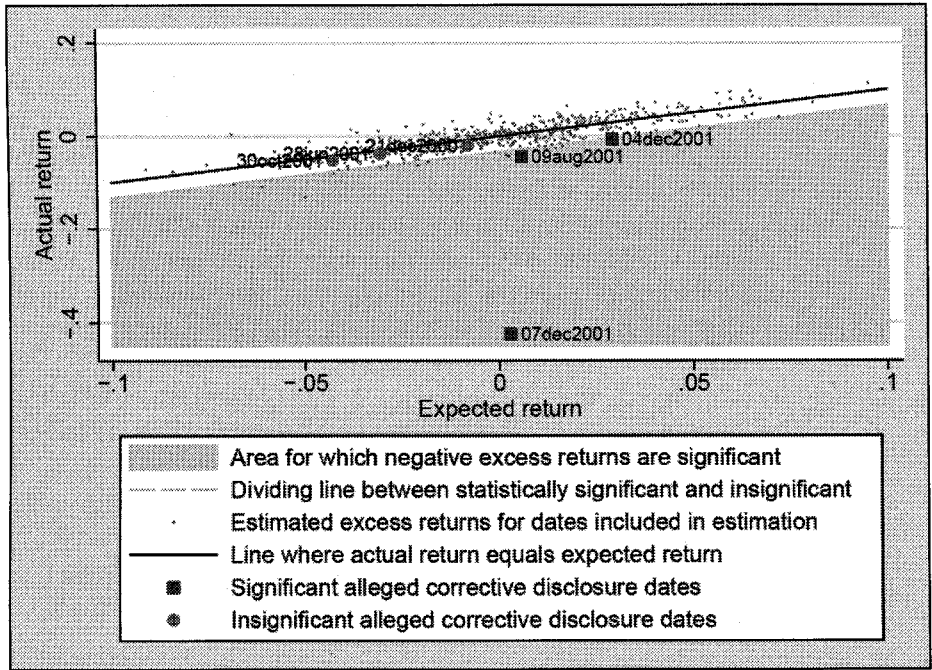
Table 4: Standard Significance Testing for Event Dates  
(sorted by magnitude of estimated excess return)

Event Date	Estimated Excess Return	Critical Value	Statistically Significant at 5 Percent Level Using Standard Approach?
December 7, 2001	-42.7%	-3.42%	Yes
August 9, 2001	-5.1%	-3.42%	Yes
December 4, 2001	-3.6%	-3.42%	Yes
December 21, 2000	-1.2%	-3.42%	No
October 30, 2001	-0.9%	-3.42%	No
June 28, 2001	-0.8%	-3.42%	No

We can illustrate the standard approach by again using a graph that relates actual and expected returns. As in earlier figures, Figure 2 again plots the actual return on the vertical axis and the expected return on the horizontal axis (with the set of points where these variables are equal indicated using an upwardly sloped straight line). This figure also includes dots indicating the expected and actual return for each day in the estimation period—these are the dots that cluster around the upwardly sloped line.



Figure 2: Scatter Plot of Actual and Expected Returns for Alleged Corrective Disclosure Dates and for Observations in Estimation Period



In addition, the figure includes three larger circles and three larger squares. The circles indicate the alleged corrective disclosure dates for December 31, 2000; October 30, 2001; and June 28, 2001—the alleged corrective disclosure dates on which Table 4 tells us estimated excess returns were negative (below the upwardly sloped line) but not statistically significant according to the standard approach. The squares indicate the alleged corrective disclosure dates for which estimated excess returns were both negative and statistically significant at the 5% level. These are the three dates in the top three rows of Table 4—December 7, 2001; August 9, 2001; and December 4, 2001. We can tell that the price drops on these dates were statistically significant at the 5% level because they appear in the shaded region of the graph; as discussed in relation to Figure 3, *infra*, points in this region have statistically significant price drops at the 5% level according to the standard approach. In sum, our implementation of a standard event study shows price impact for three dates, and it fails to show such impact at the 5% level for the other three.

#### IV. Special Features of Securities Fraud Litigation and Their Implications for the Use of Event Studies

The validity of the standard approach to testing for statistical significance, at whatever significance level is chosen, relies importantly on four assumptions:

- (1) Halliburton's excess returns actually follow a normal distribution—that assumption is the source of the 1.96 multiplier for the standard deviation of Halliburton's estimated excess returns in estimating the critical value.
- (2) It is appropriate to use a multiplier that is derived by considering what would constitute an unusual excess return in either the positive or negative direction—i.e., an unusually large unexpected movement of the stock in either the direction of increase or the direction of decrease.
- (3) It is appropriate to analyze each event date test in isolation without taking into account the fact that multiple tests (six in our *Halliburton* example) are being conducted.
- (4) Under the null hypothesis, Halliburton's excess returns have the same distribution on each date; under the first assumption (normality), this is equivalent to assuming that the standard deviation of Halliburton's excess returns is the same on every date.

As it happens, each of these assumptions is false in the context of the *Halliburton* litigation. The court did take appropriate account of the falsity of the third assumption (involving multiple comparisons),<sup>173</sup> but it failed even to address the other three.

Violations of any of these assumptions will render the standard approach to testing for statistical significance unreliable. That is true even if these violations do not always cause the standard approach to yield incorrect conclusions—i.e., conclusions that differ from what reliable methods would yield—concerning statistical significance at the chosen significance level. Just as a stopped clock is right twice a day, an unreliable statistical method will yield the right answer *sometimes*.<sup>174</sup> But the law demands more—it demands a method that yields the right answer as often as asserted by those using the method.

In the remaining sections of this Part, we explain these four assumptions in more detail, and we show that they are unsustainable in the context of the *Halliburton* event study conducted in Part III.

---

173. *Erica P. John Fund, Inc. v. Halliburton Co.*, 309 F.R.D. 251, 265–67 (N.D. Tex. 2015).

174. For example, a policy of never rejecting the null hypothesis would make no Type I errors, and a policy of always rejecting the null hypothesis would make no Type II errors. Yet both policies are obviously indefensible.



*A. The Inappropriateness of Two-Sided Tests*

In a purely academic study, economic theory may not predict whether an event date excess return can be expected to be positive or negative. For example, an announced merger might be either good or bad for a firm's market valuation. In such cases, statistical significance is appropriately tested by checking whether the estimated excess return is large in magnitude regardless of its sign. In other words, either a very large drop or a very large increase in the firm's stock price constitutes evidence against the null hypothesis that the news had no impact on stock price. Such tests are known as "two-sided" tests of statistical significance since a large value of the excess return on either side of zero provides evidence against the null hypothesis.<sup>175</sup>

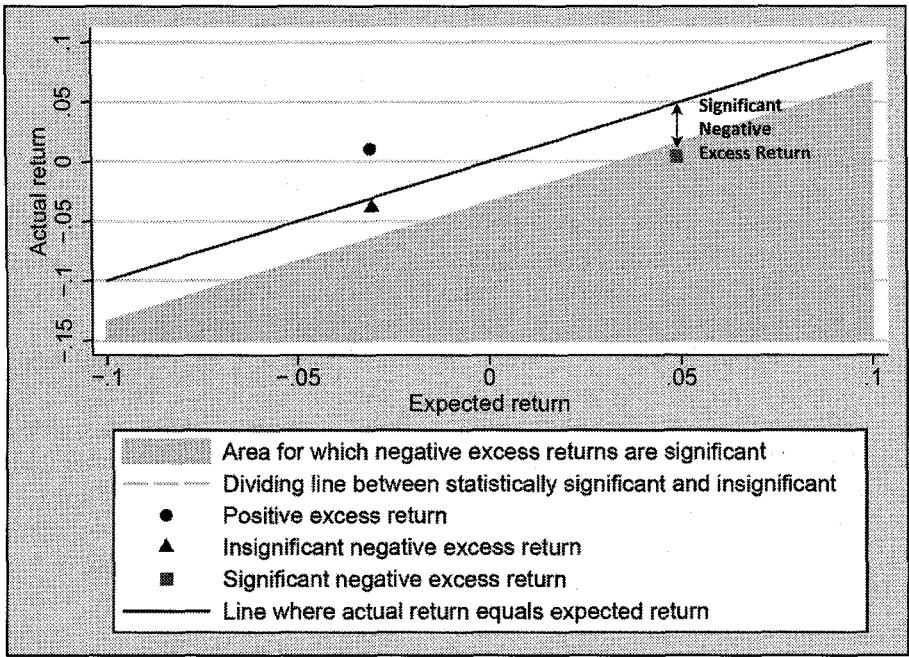
In event studies used in securities fraud litigation, by contrast, price must move in a specific direction to support the plaintiff's case. For example, an unexpected corrective disclosure should cause the stock price to fall. Thus, tests of statistical significance based on event study results should be conducted in a "one-sided" way so that an estimated excess return is considered statistically significant only if it moves in the direction consistent with the allegations of the party using the study. The one-sided–two-sided distinction is one that courts and expert witnesses regularly miss, and it is an important one.

Figure 3 illustrates this point. As in Figure 1, the upwardly sloped line indicates the set of points where the actual and excess returns are equal. The shaded area in Figure 3 depicts the set of points where the actual return is far enough below the expected return—i.e., where the excess return is sufficiently negative—so that the excess return indicates a statistically significant price drop on the date in question.

---

175. See MacKinlay, *supra* note 114, at 28 (providing an example of a two-sided test and explaining that the null hypothesis would be rejected if the abnormal return was above or below certain thresholds).

Figure 3: Illustrating Statistical Significance of Excess Returns

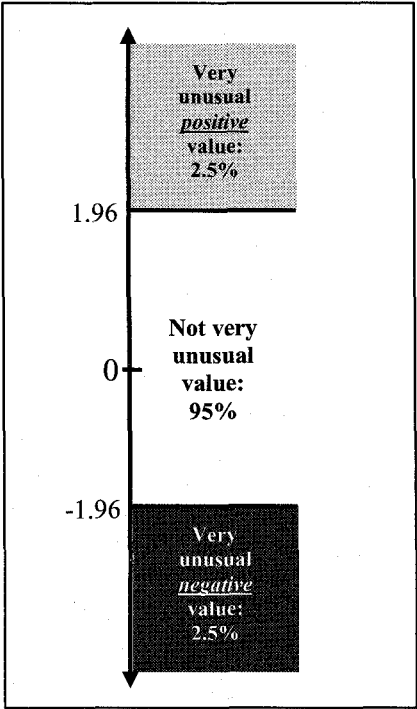


Consider the points indicated by a circle and a square in Figure 3, which are equally far from the actual-equals-expected line but in opposite directions. The circle depicts a point that has a positive excess return. Even though the circle is sufficiently far away from the line, the point has the wrong sign for an alleged corrective disclosure date, and no court would consider such evidence a basis on which to find for the plaintiff. The square, in contrast, depicts an excess return that is both negative and sufficiently far below the expected return such that we conclude there was a statistically significant price drop at the chosen significance level—as would be necessary for a plaintiff alleging a corrective disclosure. Finally, consider the point indicated by a triangle. This point is in the direction consistent with the plaintiff’s allegations—a negative excess return for an alleged corrective disclosure—but at this point the actual and expected returns are too close for the excess return to be statistically significant at the chosen level. For an alleged corrective disclosure date, only the square would provide statistically significant evidence.

If no litigant would present evidence of a statistically significant price movement in the wrong direction, why does the two-sided approach matter? The reason is that the practical effect of this approach is to reduce the Type I error rate for the tests used in event studies from the stated level of 5% to half that size, i.e., to 2.5%. To see why, consider Figure 4. Higher points in the figure correspond to larger and more positive estimated excess returns. The

shaded regions correspond to the sets of excess returns that are further from zero than the critical value of 1.96 standard deviations used by experts who deploy the two-sided approach. For each shaded region, the probability that a randomly chosen excess return will wind up in that region is 2.5%. Thus the probability an excess return will be in either region—and thus that the null hypothesis would be rejected if event study experts followed usual two-sided practice—is 5% in total, which is the desired Type I error rate.

Figure 4: The Standard Approach to Testing on an Alleged Corrective Disclosure Date with a Type I Error Rate of 5%  
(Measured in Standard Deviation Units)



However, on an alleged corrective disclosure date, the plaintiff’s allegation is that the price *fell* due to the revelation of earlier fraud. As noted, a finding that the date had an unusually large and *positive* excess return on that date would certainly not be credited to the plaintiff by the court. That is why only estimated excess returns that are large and *negative* are treated as statistically significant for proving price impact on an alleged corrective disclosure date. In other words, only estimated excess returns that are in the bottom shaded region in Figure 4 would meet the plaintiff’s burden. As we have seen, this region contains 2.5% of the probability when there is no actual

effect of the news in question.<sup>176</sup> This means that a finding of statistical significance would occur only 2.5% of the time when the null hypothesis is true—or half as frequently as the 5% rate that courts and experts say they are attempting to apply.<sup>177</sup>

Although a reduction in Type I errors is desirable with all else held equal, as we discussed in subpart II(B), *supra*, there is a trade-off between Type I and Type II error rates. As a result of this trade-off, the Type II error rate of a test rises—possibly dramatically—as the Type I error rate is reduced. This means that using a Type I error rate of 2.5% in an event study induces many more false negatives than using a Type I error rate of 5%.<sup>178</sup>

This mistake is easily corrected. Rather than base the critical value on the two-sided testing approach, one simply uses a one-sided critical value. In terms of Figure 4, that means choosing the critical value so that a randomly chosen excess return would turn up in the bottom shaded region 5% of the time, given that the news of interest actually had no impact. Still maintaining the assumption that excess returns are normally distributed, the relevant critical value is  $-1.645$  times the standard deviation of the stock's excess returns.<sup>179</sup> In our application, this yields a critical value for an event date excess return of  $-2.87\%$ ; any excess return more negative than this value will yield a finding of statistical significance.<sup>180</sup> This is a considerably less demanding critical value than the  $-3.42\%$  based on the two-sided approach. Consequently, switching to the one-sided test will correct an erroneous finding of no statistical significance at the 5% level whenever the estimated excess return is between  $-3.42\%$  and  $-2.87\%$ .

As it happens, none of the estimated excess returns in Table 4 has a value in this range, so correcting this error does not affect any of the statistical significance determinations we made in Part III for Halliburton. But that is

176. The fact that two-tailed tests are erroneous has been noted in recent literature. See Edward G. Fox, Merritt B. Fox & Ronald J. Gilson, *Economic Crisis and the Integration of Law and Finance: The Impact of Volatility Spikes*, 116 COLUM. L. REV. 325, 353 (2016) (acknowledging that the usual two-tailed test delivers a Type I error rate of only 2.5%); Fox, *supra* note 92, at 445 n.22 (same). Those authors seem to accept that courts will continue to use a method that is twice as demanding of plaintiffs as the method that courts say they require. We see no reason why courts should allow such a state of affairs to continue, especially one that is so easy to remedy.

177. A method that delivers many more false negatives than claimed surely raises important *Daubert* and FED. R. EVID. 702 concerns. See *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579, 594 (1993) (asserting that courts should consider known or potential rates of error of scientific techniques).

178. We discuss power implications of this issue in Part V.

179. This is so because a normally distributed random variable will take on a value less than  $-1.645$  times its standard deviation 5% of the time. If one were testing for statistical significance on the date of a nonconfirmatory alleged misrepresentation, one would use a critical value of  $1.645$  times the standard deviation of the excess return since a normally distributed random variable will take on a value *greater* than  $1.645$  times its standard deviation 5% of the time.

180. This critical value is the product of  $-1.645$  and the estimated standard deviation of  $1.745\%$ :  $-1.645 \times 1.745\% = -2.87\%$ .

just happenstance; had any of the estimated excess returns fallen in this range, our statistical significance conclusion would have changed. Further, Halliburton's median daily market value was \$17.6 billion over the estimation period, so the range of estimated excess returns that would have led to a switch—i.e.,  $-3.42\%$  to  $-2.87\%$ —corresponds to a range of Halliburton market value of nearly \$100 million. In other words, using the erroneous approach would, in the case of Halliburton, require a market value drop of almost \$100 million more than should be required to characterize the drop as highly unusual.

### *B. Non-Normality in Excess Returns*

Recall that, as discussed above, we characterize an excess return as highly unusual by looking at the distribution of excess returns on days when there is no news. The standard event study assumes that this distribution is normal.<sup>181</sup> There is no good reason, however, to assume that excess stock returns are actually normally distributed, and there is considerable evidence against that assumption.<sup>182</sup> Stocks' excess returns often exhibit empirical evidence of skewness, "fat tails," or both; and neither of these features would occur if excess returns were actually normal.<sup>183</sup>

In the case of Halliburton, we found strong evidence that the excess returns distribution was non-normal over the class period. Summary statistics indicate that Halliburton's excess returns exhibit negative skew: they are more likely to have positive values than negative ones. Further, the distribution has fat tails, with values far from the distribution's center than would be the case if excess returns were normally distributed. Formal statistical tests reinforce this story: Halliburton's estimated excess returns systematically fail to follow a normal distribution over the estimation period.<sup>184</sup>

181. See generally Gelbach, Helland & Klick, *supra* note 19 (discussing normal distribution).

182. For early evidence on non-normality, see Stephen J. Brown & Jerold B. Warner, *Using Daily Stock Returns: The Case of Event Studies*, 14 J. FIN. ECON. 3, 4–5 (1985). For more recent evidence in the single-firm, single-event context, see Gelbach, Helland & Klick, *supra* note 19, at 511, 534–37.

183. The existence of skewness indicates, roughly speaking, that the distribution of returns is weighted more heavily to one side of the mean than the other; the existence of fat tails—formally known as kurtosis—indicates that extreme values of the excess return are more likely in either direction than they would be under a normal distribution. See Brown & Warner, *supra* note 182, at 4, 9–10 (discussing the issues of skewness and kurtosis in the context of event studies that use daily stock-return data).

184. To test for normality, we used tests discussed by Ralph B. D'Agostino, Albert Belanger & Ralph B. D'Agostino, Jr., Commentary, *A Suggestion for Using Powerful and Informative Tests of Normality*, 44 AM. STATISTICIAN 316 (1990), and implemented by the statistical software Stata via the "sktest" command. This test rejected normality with a confidence level of 99.98%, due primarily to the distribution's excess kurtosis.

We illustrate the role of the normality assumption in Figure 5, which plots various probability density functions for excess returns. Roughly speaking, a probability density function tells us the frequency with which a given value of the excess return is observed. The probability of observing an excess return value less than, say,  $x$  is the area between the horizontal axis and the probability density function for all values less than  $x$ . The curve plotted with a solid line in the top part of Figure 5 is the familiar density function for a normal distribution (also known colloquially as a bell curve) with standard deviation equal to one. To the left of the point where the excess return is  $-1.645$ , the shaded area equals  $0.05$ ; this reflects the fact that a normal random variable will take on a value less than  $-1.645$  standard deviations 5% of the time. To put it differently, the 5th percentile of standard normal distribution is  $-1.645$ ; that is why we use this figure for the critical value to test for a price drop at a significance level of 5% when excess returns are normally distributed.

The curve plotted with a dashed line in the top part of Figure 5 is the probability density function for a different distribution. Compared to the standard normal distribution, the left-tail percentiles of this second distribution are compressed toward its center. That means fewer than 5% of this distribution's excess returns will take on a value less than  $-1.645$ ; the 5th percentile of this distribution is closer to zero, equal to roughly  $-1.36$ . Thus, when the distribution of excess returns is compressed toward zero relative to the normal distribution, we must use a more forgiving critical value—one closer to zero—to test for a significant price drop.

The bottom graph in Figure 5 again plots the standard normal distribution's probability density function with a solid line. In contrast to the top graph, the curve plotted with a dashed line now depicts a distribution of excess returns for which left-tail percentiles are splayed out compared to the normal distribution. The 5th percentile is now  $-2.35$ , so that we must use a more demanding critical value—one further from zero—to test for significance.

As this discussion illustrates, the assumption that excess returns are normally distributed is not innocuous: if the assumption is wrong, an event study analyst might use a very different critical value from the correct one.

It might seem a daunting task to determine the true distribution of the excess return. However, Gelbach, Helland, and Klick (GHK) show that under the null hypothesis that nothing unusual happened on the event date, the *estimated* excess return for a single event date will have the same statistical properties as the *actual* excess return for that date.<sup>185</sup> This result provides a

---

185. Gelbach, Helland & Klick, *supra* note 19, at 538–39. GHK actually use somewhat different notation; the estimated excess return described in the present Article is the same as GHK's  $\hat{\gamma}$  regression parameter. With this difference noted, our point about statistical properties is demonstrated in GHK's Appendix B. This result is practically useful provided that the number of



simple correction to the normality assumption: instead of using the features of the normal distribution to determine the critical value for statistical significance testing, we use the 5th percentile of the distribution of excess returns estimated using our market model.<sup>186</sup> GHK describe this percentile approach as the “SQ test” since the approach relies for its theoretical justification on the branch of theoretical statistics that concerns the behavior of sample quantiles, which, for our purposes, are simply observed percentiles.<sup>187</sup>

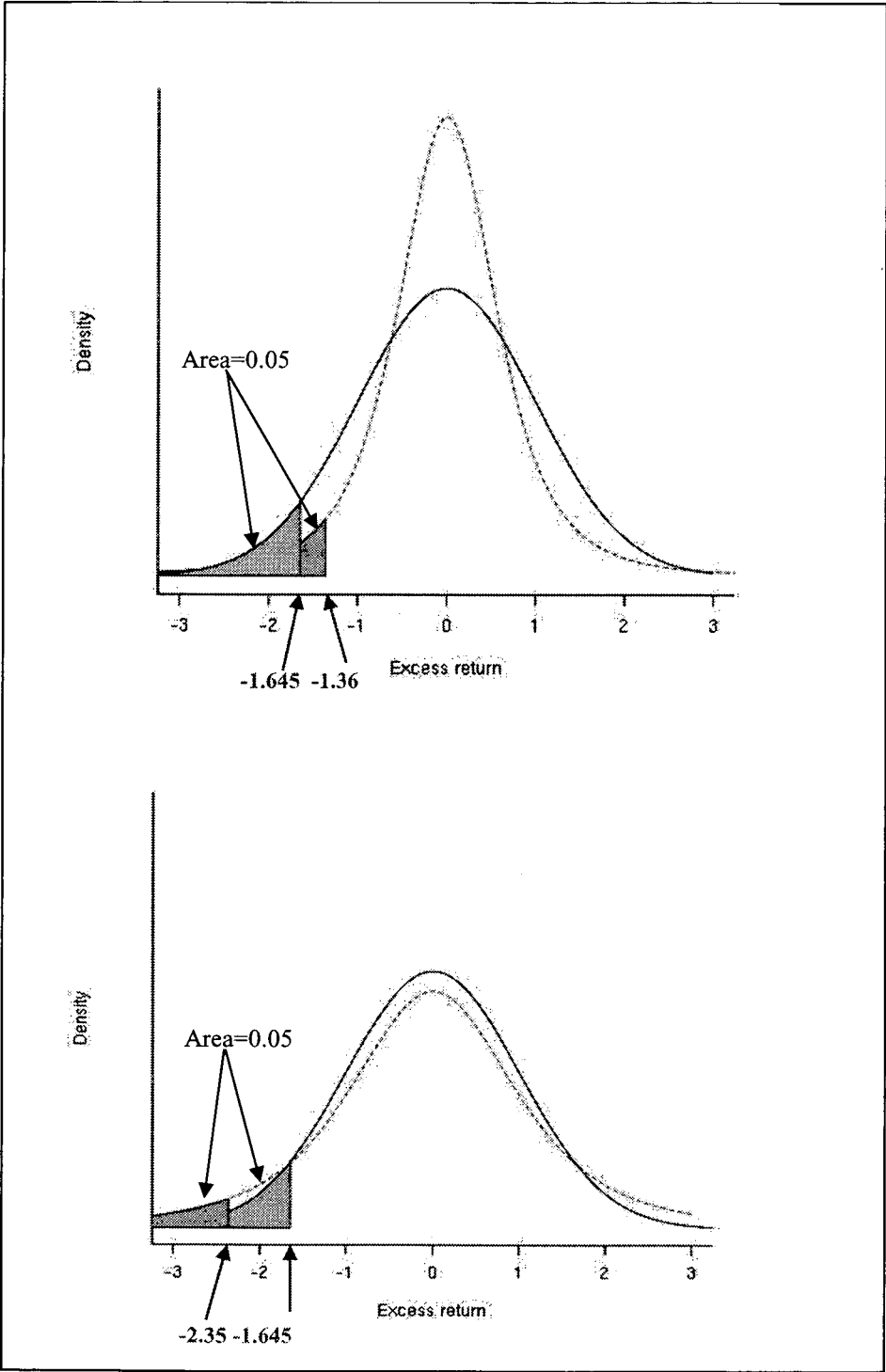
---

dates used to estimate the market model is large. We used data from July 22, 1999, through December 7, 2001, excluding the event dates at issue; this set of dates corresponds to the plaintiffs’ proposed class period at issue at the time the district court last considered class certification. *See* Erica P. John Fund, Inc. v. Halliburton Co., 309 F.R.D. 251 (N.D. Tex. 2015). This means that we used 593 dates in the market model, which is surely large in the statistically relevant sense.

186. The SQ test will erroneously reject a true null hypothesis with probability that becomes ever closer to 0.05 as the number of observations in the estimation period grows. This is an example of an asymptotic result, according to which the probability limit of the erroneous rejection probability precisely equals 0.05. Contemporary econometrics is dominated by a focus on such asymptotic results. *See, e.g.,* WILLIAM H. GREENE, *ECONOMETRIC ANALYSIS* 619 (7th ed. 2012) (discussing the absence of an asymptotic result). Unpublished tabulations from the data GHK used show that the SQ test performs extremely well even when using estimation period sample sizes considerably lower than the 250 days used here. The underlying reason the SQ test works—the reason that the standard approach’s normality assumption may be jettisoned—is that the critical value necessary for testing the null hypothesis of no event-date effect is simply the 5th percentile of the true excess returns distribution. Due to an advanced statistics result known as the Glivenko–Cantelli theorem, the percentiles of this distribution—also known as quantiles—may be appropriately estimated using the percentiles of the estimated excess returns distribution. For details, see section 5.1 of Gelbach, Helland & Klick, *supra* note 19, at 517–20.

187. Gelbach, Helland & Klick, *supra* note 19, at 497.

Figure 5: Illustrating Non-Normality





For a statistical significance test with a significance level of 5%, the SQ test entails using a critical value equal to the 5th percentile of the estimated excess returns distribution among non-event dates. Among the 593 non-event dates in our class period estimation sample, the 5th percentile is  $-3.08\%$ .<sup>188</sup> According to GHK's SQ test, then, this is the value we should use as the critical value for testing whether event date excess returns are statistically significant. Thus, when we drop the normality assumption and instead allow the distribution of estimated excess returns to drive our choice of critical values directly, we conclude that an alleged corrective disclosure date's estimated excess return is statistically significant if it is less than  $-3.08\%$ .

Note that this critical value is greater than the value of  $-2.87\%$  found in subpart IV(A), *supra*, where we maintained the assumption of normality. Thus, relaxing the normality assumption has the effect of making the standard for a finding of statistical significance about 0.21 percentage points more demanding.<sup>189</sup> Although this correction does not affect our determination as to any of the six event dates in our Halliburton event study, it is nonetheless potentially quite important because 0.21 percentage points corresponds to a range of Halliburton's market value of nearly \$40 million.

As we discuss in our online Appendix A, the SQ test has both statistical and operational characteristics that make it very desirable. First, it involves estimating the exact same market model as the standard approach does. It requires only the trivial additional step of sorting the estimated excess return values for the class period in order to find the critical value—something that statistical software packages can do in one easy step in any case. The operational demands of using the SQ test are thus minor, and we think experts and courts should adopt it. And second, the SQ test not only is appropriate in many instances where the normality assumption fails but also is always appropriate when the normality assumption is valid. Thus there is no cost to using the SQ test, by comparison to the standard approach of assuming normality.

### C. Multiple Event Dates of Interest

The approaches to statistical significance testing discussed above were all designed for situations involving the analysis of a single event date. As we have seen, however, there are six alleged corrective disclosure dates at issue in the *Halliburton* litigation. The distinction is important.

---

188. We find the 5th percentile of a sample by multiplying the number of dates in the sample by 0.05, which yields 29.65. Conventionally, this means that the 5th percentile lies between the 29th and 30th most negative estimated excess returns; in our sample, these are  $-3.089066\%$  and  $-3.074954\%$ . (The shares of estimated excess returns less than or equal to these values are 4.89% and 5.06%. Their midpoint is  $-3.08201\%$ , which is our estimate of the 5th percentile.)

189. That is, an estimated excess return must now be more negative than  $-3.08\%$ , rather than  $-2.87\%$ , to be found statistically significant.

The more tests one does while using the same critical value, the more likely it is that at least one test will yield a finding of statistical significance at the stated significance level even when there truly was no price impact. More event dates means more bites at the same apple, and the odds the apple will be eaten up increase with the number of bites. At the same time, however, securities litigation differs from the example in that multiple events do not always relate to the same fraud. Corrective disclosures relating to different misstatements are different pieces of fruit. We discuss the multiple comparison adjustment first, in section 1, and then, in section 2, we explain an approach for determining when such an adjustment is warranted. In section 3, we address the very different statistical problem raised by a situation in which a plaintiff must prove *both* the existence of price inflation on the date of an alleged misrepresentation *and* the existence of a price drop on the date of an alleged corrective disclosure.<sup>190</sup>

1. *When the question of interest is whether any disclosure had an unusual effect.*—In our event study analysis so far, we have tested for statistical significance as if each of the six event dates’ estimated excess returns constituted the only one being tested. As mentioned above, this means the probability of finding *at least* one event date’s estimated excess return significant will be considerably greater than the desired Type I error rate of 5%. The defendants raised the multiple comparison issue in the *Halliburton* litigation, and it played a substantial role in the court’s analysis.<sup>191</sup>

Various statistical approaches exist to account for multiple testing.<sup>192</sup> One approach is called the Holm–Bonferroni *p*-value correction. The district court used this approach in *Halliburton*.<sup>193</sup> To understand this correction, it is first necessary to explain the term *p*-value. The *p*-value can be viewed as another way of describing statistical significance. In terms of our prior analysis, if the estimated excess return for a single date is statistically significant at the 5% level, then the *p*-value for that date must be less than or equal to 0.05. If, on the other hand, the estimated excess return is not statistically significant, then the *p*-value must be above 0.05. We will refer to *p*-values that are computed as if only a single date were being tested as

---

190. Cases that present a combination of the questions addressed in sections 1 and 2 are more complicated notationally and mathematically; we address such cases in our online Appendix A.

191. See *Erica P. John Fund, Inc. v. Halliburton Co.*, 309 F.R.D. 251, 266 (N.D. Tex. 2015) (finding that “a multiple comparison adjustment is proper in this case”).

192. Some of them solve the Type I error rate problem at the cost of substantially increasing the Type II error probability—i.e., substantially reducing the power of the test to detect price impact where it actually occurred. As multiple testing methodology involves some fairly technical mathematical details, we will not discuss it in detail. For a brief but exceedingly clear discussion see Hervé Abdi, *Holm’s Sequential Bonferroni Procedure*, in *ENCYCLOPEDIA OF RESEARCH DESIGN* 573 (Neil J. Salkind ed., 2010).

193. *Halliburton*, 309 F.R.D. at 266–67.

“usual”  $p$ -values; this allows us to distinguish between usual and multiple-comparison-adjusted  $p$ -values.

Calculating the usual  $p$ -value for an alleged corrective disclosure date when using the one-sided SQ test involves counting up the number of estimated excess returns from the market model estimation period that are more negative than the estimated excess return on the event date and then dividing by the number of dates included when estimating the market model (593 in our *Halliburton* example). We report the usual  $p$ -value for each alleged corrective disclosure date in the second column of Table 5; the third column reports whether price impact was found statistically significant at the 5% level using the one-sided SQ test. Note that the usual  $p$ -value is less than 0.05 for all three dates with price impacts that are statistically significant at the 5% level and greater than 0.05 for the other three.

Table 5: Controlling for Multiple Testing  
Using the Holm–Šidák Approach

Event Date	Excess Return	One-Sided SQ Approach, Ignoring Multiple Testing Issue		One-Sided SQ Approach, With Šidák Correction for Multiple Testing	
		$p$ -value	Statistically Significant at 5% Level?	$p$ -value	Statistically Significant at 5% Level?
December 7, 2001	-42.7%	0	Yes	0	Yes
August 9, 2001	-5.1%	0.0017	Yes	0.0034	Yes
December 4, 2001	-3.6%	0.0269	Yes	0.0787	<b>NO</b>
December 21, 2000	-1.2%	0.2222	No	0.6340	No
October 30, 2001	-0.9%	0.2609	No	0.7795	No
June 28, 2001	-0.8%	0.3013	No	0.8837	No

The fourth column of the Table reports  $p$ -values that are corrected for multiple testing.<sup>194</sup> The final column reports whether the Holm–Šidák  $p$ -

194. There are different flavors of  $p$ -values that correct for multiple comparisons. The type we have reported in the Table is known as Šidák. Abdi, *supra* note 192, at 575. To calculate the Šidák  $p$ -value for the event date with the lowest usual  $p$ -value is just that usual  $p$ -value; thus the  $p$ -value for the excess return on December 7, 2001, is unaffected by the correction for multiple comparisons. Let the second lowest usual  $p$ -value be called  $p_2$  (December 4, 2001, in our event study). The formula for the Šidák  $p$ -value for this date is  $p_{s2} = 1 - (1 - p_2)^2$ . The logic of this formula is that the probability of independently drawing two excess returns that are more negative than the usual

value is less than 0.05, in which case there is statistically significant price impact even after adjusting for the presence of multiple tests.<sup>195</sup> Table 5 shows that after correcting for multiple testing, we find significant price impacts at the 5% level for December 7, 2001, and August 9, 2001, but not for the other four dates. Thus, relative to the one-sided SQ test that does not correct for multiple tests, the effect of correcting for multiple tests is to convert the finding of statistical significance at the 5% level for December 4, 2001, to a finding of insignificance.

2. *How should events be grouped together to adjust for multiple testing?*—A critical threshold question before applying a multiple comparison adjustment is to determine which, if any, of a plaintiff's multiple alleged corrective disclosure dates should be grouped together. In the preceding section we grouped all dates together because that is the approach the district court took in the *Halliburton* litigation.<sup>196</sup> However, it is not clear that this is the best—or even a good—approach. As noted, using multiple event dates gives the plaintiff an advantage by increasing the chance of achieving statistical significance with respect to each transaction.

How do we identify which disclosure dates to group together? A full analysis of this mixed question of law and advanced statistical methodology is beyond the scope of this Article, but one simple solution is to draw an analogy to general principles of claim preclusion. Rule 18(a)'s generous claim-joinder rule allows, but does not require, a plaintiff to bring all possible

$p$ -value actually observed on this date, i.e.,  $p_{S2}$ , is  $(1 - p_2)^2$ ; thus the probability of *not* drawing a more negative excess return is  $p_{S2}$ . The value  $p_{S2}$  is thus the probability of taking two draws from the excess returns distribution and observing at least one with a more negative excess return than  $p_2$ . It can be shown that when this probability is less than 0.05, the underlying statistic is statistically significant at the 5% level.

For the event date with the third lowest usual  $p$ -value, which we will call  $p_3$ , the formula for the Šidák  $p$ -value for this date is  $p_{S3} = [1 - (1 - p_3)^3]$ ; again the logic is that this is the probability of drawing repeatedly (now, three times) from the excess returns distribution and obtaining an excess return that is more negative than the date in question. In general, let the usual  $p$ -value for the date with the  $m^{\text{th}}$ -lowest usual  $p$ -value be  $p_m$ ; then the Šidák  $p$ -value for this date is  $p_{Sm} = 1 - (1 - p_m)^m$ . See *id.* at 576 (equation (8)). We note also that for small values of  $p_m$  and small values of the exponent  $m$ , Šidák  $p$ -values are well-approximated by  $m \times p_m$ , which is known as the Bonferroni  $p$ -value. *Id.* (equation (9)). In our application it turns out not to matter which of the two approaches we use, though in general, the Šidák  $p$ -value is more accurate than the Bonferroni  $p$ -value. *Id.* at 575–76. The district court in the *Halliburton* litigation addressed the choice between Bonferroni and Šidák  $p$ -values because experts in the case debated which was more appropriate. *Halliburton Co.*, 309 F.R.D. at 265–67. In this case, the choice makes no difference to the actual statistical significance determinations.

195. That is, we consider the price impact on the date with the second-lowest  $p$ -value to be significant only if its Šidák  $p$ -value is less than 0.05. If date six's price impact is not statistically significant, then we consider all dates' price impacts to be insignificant. If date five's price impact is significant, then we turn to considering date four's price impact, considering it significant if date four's Šidák  $p$ -value is less than 0.05; if not, we stop, but if so, we turn to date three's price impact, and so on.

196. *Halliburton*, 309 F.R.D. at 265–66.

claims in a single lawsuit.<sup>197</sup> Thus, a plaintiff might choose to bring separate actions with only a subset of alleged corrective disclosure dates at issue in each action. The rules of claim preclusion impose a limit on plaintiffs' power to litigate multiple claims independently, however, by looking to whether two claims are sufficiently closely related.<sup>198</sup> If so, a judgment on one such claim will preclude a separate cause of action on the second.

We suggest that if a losing judgment in Claim 1 would preclude a plaintiff from prevailing on Claim 2, then it is reasonable for the district court to consider all alleged corrective disclosure dates for the two claims together for purposes of multiple comparisons. Contrariwise, if losing on Claim 1 would not preclude Claim 2, then, we suggest, the alleged corrective disclosure dates related to the two claims should be treated separately. This rule would ensure that in addressing multiple alleged corrective disclosure dates, courts require a consistent quantum of statistical evidence to obtain class certification across collections of dates concerning the same or related misstatements—i.e., claims that plaintiffs would naturally be expected to litigate together. Basing this test on the law of claim preclusion prevents future plaintiffs from gaming the system by attempting to bring multiple lawsuits in order to avoid the multiple comparison adjustment. At the same time, our rule would not penalize a plaintiff for bringing two unrelated claims in the same action—thereby respecting and reinforcing the baseline set by Rule 18(a).

To illustrate with respect to *Halliburton*, five of the six alleged corrective disclosures analyzed there involved allegations related to Halliburton's asbestos liabilities.<sup>199</sup> The sixth alleged corrective disclosure date (December 21, 2000) involved Halliburton's statements regarding merger-related and other issues.<sup>200</sup> Assuming that the asbestos-related fraud allegations are sufficiently separate from the merger and other allegations that judgment in one set of claims would not preclude the other, the district court should have treated the December 21, 2000 date separately from the other five alleged corrective disclosure dates. This means that there would be no necessary correction for multiple comparisons for December 21, 2000; statistical significance testing for that date would follow the usual practice.

197. FED. R. CIV. P. 18(a) ("A party asserting a claim . . . may join, as independent or alternative claims, as many claims as it has against an opposing party.").

198. Whether the claims are closely enough related is likely to be governed by the "transaction" test. RESTATEMENT (SECOND) OF JUDGMENTS § 24 (AM. LAW INST. 1980). The Restatement is of course not *per se* binding on federal courts, but the Supreme Court has endorsed the Restatement's approach. *See, e.g.,* *United States v. Tohono O'Odham Nation*, 563 U.S. 307, 316 (2011) ("The now-accepted test in preclusion law for determining whether two suits involve the same claim or cause of action depends on factual overlap, barring 'claims arising from the same transaction.'" (quoting *Kremer v. Chem. Constr. Corp.*, 456 U.S. 461, 482 n.22 (1982), and citing RESTATEMENT (SECOND) OF JUDGMENTS § 24 (AM. LAW INST. 1980))).

199. Coffman Report, *supra* note 154, ¶ 8.

200. Allen Report, *supra* note 147, ¶ 11.

For the other five dates, the relevant number of tests would be five, rather than six as used by the district court.<sup>201</sup>

It can be shown that this change would not affect any of the statistical significance conclusions in our *Halliburton* event study. However, the change would have made a difference in other circumstances. For example, had the usual  $p$ -value for December 21, 2000, been below 0.05, it would again be considered statistically significant at the 5% level using our approach to grouping alleged corrective disclosure dates.<sup>202</sup> This example helps illustrate the importance of a court's approach to determining the number of relevant dates for purposes of adjusting for multiple testing.<sup>203</sup>

3. *When the question of interest is whether both of two event dates had an effect of known sign.*—There is another side to the multiple comparison adjustment. Consider the situation in which the plaintiff alleges that the defendant made a misrepresentation involving nonconfirmatory information on Date One and then issued a corrective disclosure on Date Two. At class certification, the plaintiff need not establish loss causation, so only price impact on Date One would be at issue. However, both dates are relevant for merits purposes since the plaintiff will have to prove both that the alleged misrepresentation caused the stock price to rise and that the alleged corrective disclosure caused the price to drop.

When the plaintiff is required to show price impact for *both* Date One *and* Date Two, the situation differs from the one considered above where it was sufficient for the plaintiff to show price impact as to *any* of multiple dates. This case is the polar opposite of that presented in the *Halliburton* litigation and requires a different statistical adjustment. In the case in which

201. It is true that this rule would require the district court to engage in a claim preclusion analysis that would otherwise be unnecessary. However, such analysis will usually not be all that cumbersome, and it provides a principled basis for determining when a multiple comparisons adjustment is appropriate. Further, the decision related to a claim preclusion question might have issue-preclusive effect, clarifying the scope of feasible subsequent litigation. That said, preclusion raises a number of serious issues in the class action setting. For a discussion, see Tobias B. Wolff, *Preclusion in Class Action Litigation*, 105 COLUM. L. REV. 717 (2005).

202. Recall from Table 5 (*supra* at 146) that the usual  $p$ -value for this date is 0.2222, whereas the  $p$ -value after correcting for multiple testing in the way the district court endorsed was 0.6340. Suppose the usual  $p$ -value had been 0.04. Then the district court-endorsed approach—treating December 21, 2000, as part of the same group as the other five dates for multiple testing purposes—would have yielded a Holm-Šidák  $p$ -value of 0.0784. Thus the district court's approach would not find statistical significance, whereas our preclusion-based approach would.

203. Still another issue that arises here involves the problem that would arise if a plaintiff's expert tested some dates but then excluded consideration of them from her expert report in order to hold down the magnitude of the multiple testing correction. *Halliburton* suggested that the plaintiffs had done just that. *Erica P. John Fund, Inc. v. Halliburton Co.*, 309 F.R.D. 251, 264 (N.D. Tex. 2015). *Halliburton* also argued that all dates on which news similar to the alleged corrective disclosures was released should be considered for purposes of determining the magnitude of the multiple testing correction. *Id.* The judge rejected the allegations of unscrupulous behavior as a factual matter. *Id.*



two events must both be shown to have statistical significance, the statistical threshold for finding price impact must be adjusted to be less demanding than if only a single date is being analyzed.

To see why, consider what would happen if we used a traditional one-sided test for each date separately, separately demanding a 5% Type I error rate for each. For each day considered in isolation, we have seen that the probability of finding statistical significance when there was no actual price impact is one in twenty. Because these significance tests are roughly independent,<sup>204</sup> the probability that *both* tests will reject when each null hypothesis is true is only one in 400, i.e., one-quarter of 1%.<sup>205</sup> To put it differently, requiring each date separately to have a 5% Type I error rate for a finding of statistical significance is equivalent to requiring a Type I error rate of just 0.25% in determining whether the plaintiff has met its merits burden as to the alleged misrepresentation in question. This is obviously a *much* more demanding standard than the 5% Type I error rate that courts and experts say they are using.<sup>206</sup>

To make an appropriate adjustment, we can again work with the usual *p*-values. For an overall *p*-value equal to 0.05—again, corresponding to the standard that experts say they are applying—we should determine that price impact is significant on both days if each date has a usual *p*-value of less than 0.2236.<sup>207</sup> Using the one-sided SQ approach, this means that the estimated price impact is statistically significant at the 5% level for the two days treated as a bundle if:

- (1) the estimated price impact for the alleged corrective disclosure date is more negative than estimated excess returns for fewer than 22.4% of the dates in the estimation period; and

204. There are two potential reasons to question independence of the estimated excess returns. First, suppose Date One involves an alleged misrepresentation and Date Two an alleged corrective disclosure. If the alleged fraud is a real one, then the magnitudes of the excess returns on Dates One and Two will be correlated. However, this fact is irrelevant to Type I error rate considerations in statistical significance testing. Such testing imposes the null hypothesis that there was actually no material fraud, in which case there is no reason to think the excess returns will be correlated. Second, though, the estimated excess returns will have a bit of dependence because they are calculated from the same estimated market model for which estimated coefficients will be common to the two event date excess returns. However, this dependence can be shown to vanish as the number of dates in the estimation period grows, and with 593 dates we would expect very little to persist.

205. This is the case because  $1/20$  times itself is  $1/400$ , which is one-fourth of  $1/100$ —or, equivalently, a quarter of a percent.

206. In terms of confidence level, the actual standard amounts to 99.75% confidence rather than the claimed 95%.

207. This is true because the probability of finding that two independent tests have a usual *p*-value of  $q$  is  $q^2$ . Setting this equal to 0.05 and solving for  $q$  yields  $q = 0.2236068$ . Thus, we should declare the *pair* of price impact estimates jointly significant if each has a usual *p*-value less than this level.

- (2) the estimated price impact for the alleged misrepresentation date is greater than estimated excess returns for fewer than 22.4% of the dates in the estimation period.

The resulting test has a 5% Type I error rate, i.e., a 5% chance of erroneously making a finding of statistical significance as to both dates considered together.

To illustrate using our *Halliburton* example, think of December 21, 2000, as Date Two, and imagine that the alleged corrective disclosure on that date had been associated not with a confirmatory disclosure but a nonconfirmatory alleged misrepresentation on Date One. In that case, the plaintiff would have to prove both that the stock price rose an unusual amount on Date One *and* that it fell by an unusual amount following the alleged corrective disclosure on December 21, 2000. Recall that the usual *p*-value for the December 21, 2000 estimated excess return was 0.2222.<sup>208</sup> This value just makes the 0.2236 cutoff. If the hypothetical Date One estimated excess return had a usual *p*-value of 0.2236 or lower, then both arms of our test would be met.

In such a case, a court using the 5% significance level should find that the plaintiff carried its burden to show both a material change in price for the alleged misrepresentation and loss causation as to the alleged corrective disclosure on December 21, 2000. This conclusion follows *even though we would not find statistically significant evidence of price impact at the 5% level if December 21, 2000, were the only date of interest*. This example illustrates the consequences of the appropriate loosening of the threshold for finding statistical significance when a party must demonstrate that something unusual happened on each of multiple dates.

We know of no case where our argument has even been made, but it is grounded in the same statistical analysis applied by the court in *Halliburton*. Concededly, a court could take the view that for any single piece of statistical evidence to be credited, that single piece must meet the 5% Type I error rate—even if that means that a party who must show two pieces of evidence is actually held to the radically more demanding standard of a 0.25% Type I error rate.<sup>209</sup> We believe that such a view is indefensible on probability grounds.

#### *D. Dynamic Evolution of the Excess Return's Standard Deviation*

For a traditional event study to be probative, the behavior of the stock in question must be stable over the market model's estimation period. For

208. See *supra* Table 5.

209. We note that our point is especially important for those situations in which there are more than just two dates in question. For example, if there were five dates, then the true Type I error rate when a court requires the plaintiff to meet the 5% Type I error rate separately for each date would be less than 0.00003% (which is approximately 1 in 3.2 million—or 1/20 raised to the fifth power).



example, it must be true that, aside from the alleged fraud-related events under study, the association between Halliburton's stock and the broader market during the class period is similar to the relationship for the estimation period. If, for example, Halliburton's association with its industry peers or other firms in the broader market differed substantially in the two periods, then the market model would not be a reliable tool for predicting the performance of Halliburton's stock on event dates, even in the absence of any actual misrepresentations or corrective disclosures.

A second requirement is that, aside from any effects of the alleged misrepresentations or corrective disclosures, excess returns on event dates must have the same probability distribution as they do during the estimation period. As we discussed in subpart IV(A), *supra*, the standard approach to estimating the critical value for use in statistical significance testing is based on the assumption that, aside from the effects of any fraud or corrective disclosure, all excess returns come from a normal distribution with the same standard deviation. But imagine that the date of an alleged corrective disclosure happens to occur during a time of unusually high volatility in the firm's stock price—say, due to a spike in market uncertainty about demand in the firm's principle industry. In that case, even typical excess returns will be unusually dispersed—and thus unusually likely to fall far from zero. Failing to account for this fact would lead an event study to find statistically significant price impact on too many dates, regardless of the significance level, simply due to the increase in volatility.<sup>210</sup>

Consider an extreme example to illustrate. Suppose that the standard deviation of a stock's excess return is usually 1%, and for simplicity, assume that the excess returns always have a normal distribution. An expert who assumes the standard deviation is 1% on an alleged corrective disclosure date therefore will determine that the excess return for that date is statistically significant at the 5% level if it is less than  $-1.645\%$ .<sup>211</sup> But suppose that on the date of the alleged corrective disclosure, market uncertainty causes the firm's standard deviation to be much greater than usual—e.g., 2%. Then the actual Type I error rate for the expert's test of statistical significance is about 21%—more than four times the chosen significance level.<sup>212</sup> What has

---

210. See Allen Report, *supra* note 147, ¶¶ 229–31, 233, 236 (illustrating how market forces can impact a company's stock volatility); Fox, Fox & Gilson, *supra* note 176, at 357 (indicating that volatility can cause increased rates of statistically significant errors); Andrew C. Baker, Note, *Single-Firm Event Studies, Securities Fraud, and Financial Crisis: Problems of Inference*, 68 STAN. L. REV. 1207, 1250–51 (2016) (same).

211. Recall that for a normally distributed random variable, which has mean zero and standard deviation one, the probability of taking on a value less than  $-1.645$  is 0.05, i.e., 5%.

212. It is a fact of probability theory that the probability that a normally distributed random variable with standard deviation  $\sigma$  takes on a value less than  $-1.645$  is the same as the probability that a normally distributed random variable with standard deviation of one takes on a value less than  $-1.645/\sigma$ . Setting  $\sigma$  equal to two, the resultant probability is 0.2054, or roughly 21%.

happened here is that the increase in the standard deviation on the alleged corrective disclosure date means that the excess return is more likely to take on values further from the average of zero. Consequently, the excess return on this date is more likely than usual to correspond to a price drop of more than 1.645%. The opposite result would occur if the standard deviation were *lower* on the alleged corrective disclosure date. With a standard deviation of only one-half on that date, the Type I error rate would fall to 0.05%, which is one one-hundredth of the chosen significance level.<sup>213</sup> Ignoring the alleged corrective disclosure date's difference in standard deviation in this situation would make false negatives (Type II errors) much more common than would a test that uses a correct critical value for the alleged corrective disclosure date excess return.

Changes in volatility are a potentially serious concern in at least some cases. Fox, Fox, and Gilson show that the stock market has experienced volatility spikes in connection with every major economic downturn from 1925 to 2010, including the 2008 financial crisis.<sup>214</sup> As they point out, the effect of a volatility spike is to raise the necessary threshold for demonstrating materiality or price impact with an event study, thereby increasing the Type II error rate of standard event study tests.<sup>215</sup>

Event studies can be adjusted to deal with the problem of dynamic changes in standard deviation. To do so, one must use a model that is capable of estimating the standard deviation of the event date excess return both for dates used in the estimation period—our “usual” dates from above—and for those dates that are the object of the price impact inquiry. The details of doing so are fairly involved, requiring both a substantial amount of mathematical notation and a discussion of some technical econometric issues. Accordingly, we relegate these details to our online Appendix C, which appears at the end of this Article, and provide only a brief conceptual summary here. We use a statistical model that allows the standard deviation of excess returns to vary on a day-to-day basis—whether due to the evolution of market- or industry-level return volatility or to the evolution of Halliburton's own return volatility. To compute the *p*-value for each event date, we use the model's estimates to rescale the excess returns for non-event dates so that all these dates have the same standard deviation as each event date in question. We then use the rescaled excess returns to conduct one-sided SQ tests with correction for multiple testing, as discussed in the sections above.

---

213. Setting  $\sigma$  equal to 0.5, the probability in question is the probability that a normally distributed random variable with standard deviation of one takes on a value less than  $-3.29$ , which is 0.0005, or roughly 0.05%.

214. Fox, Fox & Gilson, *supra* note 176, at 335–36.

215. *See id.* at 357 (stating that a volatility spike “can result in a several-fold increase in Type II error—that is, securities fraud claims will fail when they should have succeeded”).

Using the approach detailed in our online appendix, we find that the standard deviation in Halliburton's excess returns does not remain stable but rather evolves over our time period in at least three important ways. First, Halliburton's excess returns have greater standard deviation on days when the industry peer index returns have greater standard deviation. Second, Halliburton's excess returns are more variable on days when a measure of overall stock market volatility suggests this volatility is greater.<sup>216</sup> Third, the standard deviation in Halliburton's excess returns tended to be greater on days when it was greater the day before and when Halliburton's actual excess return was further from zero (whether positive or negative).

Using the model estimates described in our online Appendix A, we tested for normality of the rescaled excess returns.<sup>217</sup> We found that the data resoundingly reject the null hypothesis that the white noise term  $u_t$  is distributed normally.<sup>218</sup> Accordingly, it is unreliable to base a test for statistical significance on the assumption that  $u_t$  follows a normal distribution.<sup>219</sup> We therefore use the SQ test approach described in subpart IV(B), *supra*. Table 6 reports  $p$ -values from our earlier and new results. The first three columns involve what we have called "usual"  $p$ -values, which are computed as if statistical significance were being tested one date at a time (i.e., ignoring the multiple-testing issue). The first column of these three reports the usual  $p$ -values from Table 5, which were computed from statistical significance tests that impose the assumption that the standard deviation of Halliburton's excess returns is the same on all dates. The second

---

216. This market-level measure is known as the VIX and is published by the Chicago Board Options Exchange. It uses data on options prices, together with certain assumptions about the behavior of securities prices, to back out an estimate of the variance of stock returns for the day in question. Its use as a variance forecasting tool has recently been advocated in Baker, *supra* note 210, at 1239, following such use of an event study in a securities fraud litigation. See Expert Report of Mukesh Bajaj ¶¶ 85, 88, 89 & n.150, *In re* Fed. Home Loan Mortg. Corp. (Freddie Mac) Sec. Litig., 281 F.R.D. 174 (S.D.N.Y. 2012) (No. 1:09-MD-2072 (MGC)) (cited in Baker, *supra* note 210, at 1245 n.217). We discuss Baker's approach, and its implicit assumption that standardized excess returns are normally distributed, in our online Appendix A. Finally, we note that another recent paper suggests that when the assumptions about the behavior of securities prices, referred to above, are incorrect, the VIX index does not directly measure the variance of the market return. See K. Victor Chow, Wanjun Jiang & Jingrui Li, Does VIX Truly Measure Return Volatility? 2–3 (Aug. 30, 2014) (unpublished manuscript), <http://ssrn.com/abstract=2489345> [<https://perma.cc/82WX-CPSW>] (explaining that the VIX index reliably measures the variance of the stock market only under certain assumptions and offering a generalized alternative for use in its place). Because our mission here is illustrative only, however, there is no harm in using the VIX index itself; we note in addition that the VIX index is much less important in explaining the variance of Halliburton's excess returns than is volatility in the industry peer index.

217. We used the same method as in subpart IV(B). See *supra* note 187.

218. While there is a bit of negative skew in the standardized estimated excess return, the test rejects normality primarily because of excess kurtosis—i.e., fat tails—in the standardized excess return distribution.

219. Baker appears to have done exactly this in his simulation study. See Baker, *supra* note 210, at 1246 (referring to the use of  $t$ -statistics to determine rejection rates).

column reports usual  $p$ -values computed from our model that allows the standard deviation to evolve over time. Our third column shows that when we ignore the issue of multiple tests, our conclusions from statistical significance testing are the same whether we account for dynamics in the daily standard deviation or not. (Three of the dates are found significant at the 5% level using both approaches, and the other three are not.)

The last three columns of Table 6 provide  $p$ -value and significance testing results when we take into account the fact that there are six alleged corrective disclosure dates.<sup>220</sup> For five of the six dates, the significance conclusion is unaffected by allowing Halliburton’s excess return standard deviation to vary over time. However, for December 4, 2001, the  $p$ -value drops substantially once we account for the possibility of evolving standard deviation: it falls from 0.0787, which is noticeably above the significance threshold of 0.05, to 0.03, which is almost as far below the threshold. Allowing for the evolution of standard deviation thus would have mattered critically in *Halliburton*, given that the court did account for the multiple dates on which alleged corrective disclosures must be assessed statistically.

Table 6: Controlling for Evolution in the Volatility of  
Halliburton’s Excess Returns

Event Date	Usual $p$ -Value (No Accounting for Multiple Tests)			Holm-Šidák $p$ -Value (Accounting for Six Tests)		
	Assuming Constant Standard Deviation	Allowing Dynamic Standard Deviation	Statistical Significance	Assuming Constant Standard Deviation	Allowing Dynamic Standard Deviation	Statistical Significance
December 7, 2001	0	0	Both	0	0	Both
August 9, 2001	0.0017	0.002	Both	0.0034	.003	Both
December 4, 2001	0.0269	0.010	Both	0.0787	.030	<b>Dynamic Only</b>
December 21, 2000	0.2222	0.256	Neither	0.6340	.694	Neither
October 30, 2001	0.2609	0.317	Neither	0.7795	.881	Neither
June 28, 2001	0.3013	0.298	Neither	0.8837	.851	Neither

220. See *supra* notes 194–95 (discussing the Holm–Šidák approach).

What drives this important reversal for the December 4, 2001 alleged corrective disclosure? For that date, our volatility model yields an estimated standard deviation of 1.5%. This is lower than the value of 1.745% in the constant-variance model underlying Table 5, and that is part of the story. But there is more to it. When we assumed constant variance across dates, there were sixteen estimation period dates that had a more negative estimated excess return than the one for December 4, 2001. Once we allowed for the standard deviation to evolve over time, *all but one of these sixteen dates* had an estimated standard deviation greater than 1.5%. In some cases, the difference was quite substantial, and this is what is driving the very large change in the  $p$ -value for December 4, 2001.<sup>221</sup>

In sum, the standard deviation on December 4, 2001, was a bit on the low side, while dates in the left tail of the excess returns distribution had very high standard deviations. When we multiply by the scale factor to make all other dates comparable to December 4, 2001, the rescale excess returns for left-tail dates move toward the middle of the distribution. This result indicates that the December 4, 2001 excess return is considerably more unusual than it appears when we fail to account for dynamic evolution in the standard deviation. Once we correct that failure, we find that the excess return on the alleged corrective disclosure date of December 4, 2001, is statistically significant at the 5% level.

#### *E. Summary and Comparison to the District Court's Class Certification Order*

Our analysis in this Part raises four issues that are often not addressed in event studies used in securities litigation: the inappropriateness of two-sided testing, the non-normality of excess returns, multiple-inference issues that arise when multiple dates are at issue, and dynamic volatility in excess returns. After accounting for all four of these issues in our event study using data from the *Halliburton* litigation, we find that at the 5% level there is statistically significant evidence of negative excess returns on three dates: December 7, 2001; August 9, 2001; and December 4, 2001. The district court

---

221. For example, five of the sixteen dates had estimated values of  $\sigma_t$  in excess of 0.023. While this might not seem like much of a difference, it is, because the standardized estimated excess return  $u_t$  is the ratio of the estimated excess return  $\varepsilon_t$  to the estimate of  $\sigma_t$ . Dividing the December 4, 2001 estimated excess return by 0.015 while dividing these other five dates' estimated excess returns by 0.023 is the same as increasing the December 4, 2001 estimated excess return by a factor of more than 50%. To see this, observe that since  $u_t = \frac{\varepsilon_t}{\sigma_t}$ , we have  $\frac{\varepsilon_{4Dec2001}}{\varepsilon_t} = \frac{u_{4Dec2001}}{u_t} \times \frac{0.023}{0.015} = 1.53 \times \frac{u_{4Dec2001}}{u_t}$ , so that this constellation of estimated values of  $\sigma_t$  makes a very large difference in the relative value of the December 4, 2001 alleged corrective disclosure date's standardized estimated excess return, by comparison to dates with very negative nonstandardized estimated excess returns.

certified a class related to December 7, 2001, in line with one of our results. However, it declined to certify a class with respect to the other dates.

As to August 9, 2001, the court did find that “there was a price movement on that date,”<sup>222</sup> which is in line with our statistical results. However, the court found that Halliburton had proved (i) that the information the plaintiff alleged constituted a corrective disclosure had been disclosed less than a month earlier, and that (ii) there had been no statistically significant change in Halliburton’s stock price on the earlier date.<sup>223</sup> Thus, the court found for purposes of class certification that the alleged corrective disclosure on August 9, 2001, did not warrant the *Basic* presumption.<sup>224</sup> We express no opinion as to this determination.

The court’s decision not to certify a class as to December 4, 2001, was founded entirely on its statistical findings of fact.<sup>225</sup> The court came to this finding by adopting the event study methodology used by Halliburton’s expert.<sup>226</sup> While that expert did correct for multiple inferences, she failed to appropriately deal with the other three issues we have raised in this Part. A court that adopted our methodology and findings while using the 5% level would have certified a class as to December 4, 2001. The court’s decision not to certify a class as to December 4, 2001, appears to be founded on event study evidence plagued by methodological flaws.

## V. Evidentiary Challenges to the Use of Event Studies in Securities Litigation

The foregoing Parts have explained the role and methodology of event studies and identified several adjustments required to make the event study methodology reliable for addressing issues of price impact, materiality, loss causation, and damages in securities fraud litigation. We turn, in this Part, to the limitations of event studies—what they can and cannot prove. Although event studies became popular because of the apparent scientific rigor that they bring to analysis of the relationship between disclosures and stock price movements, the question that they answer is not identical to the underlying legal questions for which they are offered as evidence. In addition, characteristics of real world disclosures may limit the ability of an event study to determine the relationship between a specific disclosure and stock

---

222. *Erica P. John Fund, Inc. v. Halliburton Co.*, 309 F.R.D. 251, 272 (N.D. Tex. 2015).

223. *Id.* at 272–73.

224. *Id.* at 273.

225. *Id.* at 276 (“[T]he Court will look only at whether there was a statistically significant price reaction on December 4, 2001.”).

226. *Id.* (“If [Halliburton’s expert’s methodology is] applied to [the plaintiff’s expert’s] model, there was no statistically significant price reaction on December 4.”). The court noted that it “ha[d] already explained that these adjustments [were] appropriate.” *Id.* It therefore found “a lack of price impact on December 4, 2001, and [that] Halliburton ha[d] met its burden of rebutting the *Basic* presumption with respect to the corrective disclosure made on that date.” *Id.*



price. Using demanding significance levels such as 5% also raises serious questions about whether statistical and legal standards of proof conflict. Finally, using event study methodology with a significance level of 5% incorporates an implicit normative judgment about the relative importance of Type I and Type II errors that masks an underlying policy judgment about the social value of securities fraud litigation. These concerns have not received sufficient attention by the courts that are using event studies to decide securities cases.

#### *A. The Significance of Insignificance*

As commonly used by scholars, event studies answer a very specific type of question: Was the stock price movement on the event date highly unusual? More precisely, event studies ask whether it would have been very unlikely to observe the excess return on the event date in the absence of some unusual firm-specific event. In the case of a securities fraud event study, the firm-specific event is a fraudulent statement or a corrective disclosure.

Importantly, event study evidence of a highly unusual excess return rebuts the null hypothesis of no price effect. But failure to rebut the null hypothesis does not necessarily mean that a misrepresentation had no price impact. An event date's excess returns might be in the direction consistent with the plaintiff's allegations but be too small to be statistically significant at a significance level as demanding as 5%. Failure to demonstrate this level of statistical insignificance does not prove the null hypothesis, however; rather, such failure simply implies that one does not reject the null hypothesis at that significance level. That is, the standard event study does not show that the information did not affect stock price; it just shows that the information did not have a statistically significant effect at the 5% level.<sup>227</sup>

This limitation raises several concerns. One is the appropriate legal standard of proof when event study evidence is involved. To our knowledge, the practice of requiring statistical significance at the 5% level at summary judgment or trial has never been justified in terms of the applicable legal standards of proof. These legal standards and the standard of statistical significance at the 5% level may well not be consistent with each other. Statistical significance concerns the unlikeliness of observing evidence if the null hypothesis of no price impact is true, whereas legal standards for adjudicating the merits are concerned with whether the null hypothesis is more likely true or false. The implications of these observations are a subject for future work.<sup>228</sup>

---

227. See Brav & Heaton, *supra* note 11, at 587 (“Courts err because of their mistaken premise that statistical insignificance indicates the probable absence of a price impact.”).

228. See generally Burtis, Gelbach & Kobayashi, *supra* note 124, at 1–3 (discussing the general mismatch between legal standards and the statistical significance testing with a fixed significance level).



A second concern is which party bears the burden of proof (whatever it is). As Merritt Fox has explained, an open issue following the Supreme Court's decision in *Halliburton II* concerns the appropriate burden of proof for a defendant seeking to rebut a plaintiff's showing of price impact at the class certification stage.<sup>229</sup> If courts continue to regard the 5% level as the right one for event studies, this distinction may be largely cosmetic. To the extent that the plaintiff will have the burden of proof at summary judgment or at trial to establish materiality, reliance, and causation, a plaintiff will need to offer an event study that demonstrates a highly unusual price effect at that time. In that case, the practical effects of imposing the burden of proof on the defendant will be short-lived.<sup>230</sup>

This in turn introduces the third concern. To what extent should courts consider additional evidence of price impact in a case in which even a well-constructed event study is unlikely or unable to reject the null hypothesis? We consider this question in more detail in subparts B and C below.

#### *B. Dealing with Multiple Pieces of News on an Event Date*

There are at least two additional ways in which the question answered by an event study differs from the legally relevant question. First, event studies cannot determine whether the event in question *caused* the highly unusual excess return.<sup>231</sup> It is possible that (i) the stock did move an unusual amount on the date in question but that (ii) some factor other than the event in question was the cause of that move. For example, suppose that on the same day that Halliburton made an alleged corrective disclosure, one of its major customers announced for the first time that it was terminating activity in one of the regions where it uses Halliburton's services. The customer's statement, rather than Halliburton's corrective disclosure, might be the cause of a drop in stock price.

Second, it is possible that the event in question *did* cause a change in stock price in the hypothesized direction, even when the estimated excess return on the event date of interest was not particularly unusual because some other factor operated in the opposite direction. For an example of this situation, suppose that Halliburton made an alleged corrective disclosure on

229. See Fox, *supra* note 92, at 438.

230. We note that deferring the dismissal of a case to, say, summary judgment would create some settlement value since both the prospect of summary judgment and the battle over class certification involve litigation costs. We leave for another day a full discussion of the importance of these costs in the long-running debate over the empirical importance of procedure in generating the filing of low-merit cases.

231. Even if the event study were capable of identifying causality, it would not be able to specifically determine the reasons for the causal reaction. Thus, as noted above, the correct response to Justice Alito's question at oral argument in *Halliburton II*, see Transcript of Oral Argument at 24, *Halliburton Co. v. Erica P. John Fund, Inc. (Halliburton II)*, 134 S. Ct. 2398 (2014) (No. 13-317), is that, by themselves, event studies are incapable of distinguishing between a rational and irrational response to information.

the same date that a major customer announced *good* news for the company. It is possible that customer's announcement would fully or partially offset the effect of the corrective disclosure, at least within the limits of the power that appropriate statistical tests can provide. In that case, there will be no highly unusual change in Halliburton's stock price—no unusual estimated excess return—even though the corrective disclosure reduced Halliburton's stock price *ex hypothesi*.

Both of these problems arise because an additional event occurs at the same time as the legally relevant alleged event. We might term this additional event a confounding event.<sup>232</sup> If multiple unusual events—events that would affect the stock price even aside from any industry-wide or idiosyncratic developments—occurred on the event date, then even an event study that controls for market- or industry-level factors will be problematic. Suppose our firm announced both favorable restructuring news and a big jury verdict against it on the same day. All a traditional event study can measure is the net market response to these two developments. Without further refinement, it would not distinguish the sources of this response.

The event study methodology might be refined to deal with some possible confounding events. For example, if the two pieces of information were announced at different times on the same day, one might be able to use intraday price changes to parse the separate impacts of the two events.<sup>233</sup> Here both the theory of and empirical evidence related to financial economics are especially important. The theory suggests that stock prices should respond rapidly in a public market with many traders paying attention to a well-known firm with many shares outstanding. After all, no one wants to be left holding a bag of bad news, and everyone can be expected to want to buy a stock for which the issuer's good news has yet to be reflected in price. These standard market factors can be expected to put immediate pressure on a firm's stock price to move up in response to good news and down in response to bad news. Empirical evidence suggests that financial economics theory is correct on this point: one widely cited, if dated, study indicates that prices react within just a few minutes to public news related to stock earnings and dividends.<sup>234</sup> As a

---

232. See, e.g., *Sherman v. Bear Stearns Cos. (In re Bear Stearns Cos., Sec., Derivative, & ERISA Litig.)*, No. 09 Civ. 8161 (RWS), 2016 U.S. Dist. LEXIS 97784, at \*28 (S.D.N.Y. 2016) (discussing whether an event study controlled sufficiently for “confounding factors”).

233. See Brav & Heaton, *supra* note 11, at 607 (discussing intraday event studies and citing *In re Novatel Wireless Sec. Litig.*, 910 F. Supp. 2d 1209, 1218–21 (S.D. Cal. 2012), in which the court held that an expert's testimony as to such a study was admissible).

234. James M. Patell & Mark A. Wolfson, *The Intraday Speed of Adjustment of Stock Prices to Earnings and Dividend Announcements*, 13 J. FIN. ECON. 223, 249–50 (1984). This study is cited, for example, in the report of Halliburton's expert witness Lucy Allen. Allen Report, *supra* note 147, ¶ 86 n.93. We note that if two pieces of news are released very close in time to each other, that might raise special challenges related to the limited amount of trading typically seen in a short enough window; this issue is beyond the scope of the present Article.

result, a study that looks at price movements during the day may be able to separate out the effect of disclosures that took place at different times.

When multiple sources of news are released at exactly the same time, however, no event study can by itself separate out the effects of the different news. The event study can only tell us whether the net effect of all the news was associated with an unusually large price drop or rise.

The results of the event study could still be useful if there is some way to disentangle the expected effects of different types of news. For example, suppose that a firm announces bad regulatory news on the same day that it announces bad earnings news, with plaintiffs alleging only that the regulatory news constitutes a corrective disclosure. Experts might be able to use historical price and earnings data for the firm to estimate the relationship between earnings news and the firm's stock price. If this study controlled appropriately for market expectations concerning the firm's earnings (say, using analysts' predictions), it might provide a plausible way to separate out the component of the event date's estimated excess return that could reasonably be attributed to the earnings news, with the rest being due to the alleged corrective disclosure related to regulatory news. Alternatively, experts might use quantitative content analysis, e.g., measuring the relative frequencies of two types of news in headlines of articles published following the news.<sup>235</sup> While the release of multiple pieces of news on the same date complicates the use of event studies to measure price impact, event studies might be useful in at least some of those cases. On the other hand, as this discussion suggests, an event study is likely to be incapable of definitively resolving the question of price impact, and a court considering a case involving confounding disclosures will have to determine the role of other evidence in addressing the question.

Lurking in the shadows of this discussion is the question of *why* information events might occur at the same time in a way that would complicate the use of an event study. Although the presence of confounding events could result from random chance, it could also be that an executive shrewdly decides to release multiple pieces of information simultaneously.<sup>236</sup> Specifically, judicial reliance on event studies creates an incentive for issuers and corporate officials to bundle corrective disclosures with other information in a single press release or filing. If the presence of overlapping news makes it difficult or impossible for plaintiffs to marshal admissible and useful event study evidence, defendants may strategically structure their disclosures to impede plaintiffs' ability to establish price effect. The

---

235. TABAK, *supra* note 10, at 13 (discussing a hypothetical scenario where the importance of different news stories can be distinguished quantitatively).

236. There is some evidence that corporate officials are able to reduce the cost of securities litigation through the use of information bundling. Barbara A. Bliss, Frank Partnoy & Michael Furchtgott, *Information Bundling and Securities Litigation* 2–4 (San Diego Legal Studies, Paper No. 16-219, 2016), <https://ssrn.com/abstract=2795164> [<https://perma.cc/9UJU-R54J>].

possibility of such strategic behavior raises important questions about the admissibility of non-event study evidence.

*C. Power and Type II Error Rates in Event Studies Used in Securities Fraud Litigation*

The focus of courts and experts in evaluating event studies has been on whether an event study establishes a statistically significant price impact at the 5% level. As we discussed briefly in regard to Table 1 in subpart II(B), *supra*, the 5% significance level requires that the Type I error rate be less than 5%. But Type I errors are only one of two ways an event study can lead to an erroneous inference. An event study leads to a Type II (false negative) error when it fails to reject a null hypothesis that really is false—i.e., when it fails to detect something unusual that really did happen on a date of interest.

As we discussed in subpart II(B), *supra*, for a given statistical test there is a trade-off between Type I and Type II error rates—choosing to tolerate fewer false positives necessarily creates more false negatives. Thus, by insisting on a 5% Type I error rate, courts are implicitly insisting on both a 5% rate of false positives and some particular rate of false negatives. Recent work has pointed out that in single-firm event studies used in securities litigation, requiring a Type I error rate of only 5% yields an extremely high Type II error rate.<sup>237</sup>

To illustrate, suppose that a corrective disclosure by an issuer actually causes a price drop of 2%. We assume for simplicity that the issuer's excess returns are normally distributed with a standard deviation of 2%.<sup>238</sup> A properly executed event study that uses the 5% level will reject the null hypothesis of no effect on that date only if the estimated excess return represents a price drop of more than -1.645%. The probability that this will occur when the true price effect is 2%—also known as the power of the test against the specific alternative of a 2% true effect—is 57%.<sup>239</sup> This means that the Type II error rate is 43%.<sup>240</sup> In other words, 43% of the time, the

---

237. See Brav & Heaton, *supra* note 11, at 597 (discussing the fact that the Type II error rate is 73.4% for a stock with normally distributed excess returns having a standard deviation of 1.5%, when the true event-related price impact is a drop of 2%).

238. This magnitude for the standard deviation was not atypical in 2014. See, e.g., Brav & Heaton, *supra* note 11, at 595 tbl.1 (showing that the average value of the standard deviation of excess returns was 2% among firms for which standard deviations put them in the sixth decile of 4,298 firms studied for 2014).

239. Because the excess return is assumed normally distributed with standard deviation 2%, the scaled random variable that equals one-half the excess return will have a normal distribution with mean zero and standard deviation 1%. Since the corrective disclosure causes a 2% drop, the event study described in the text will yield a finding of statistical significance whenever -1 plus this scaled random variable is less than the ratio (-1.645/2). The probability of that event—the test's power in this case—can be shown to equal 0.5704.

240. Since the probability of a Type II error is one minus the power of the test, the probability of a Type II error is 0.4296, which implies a Type II error rate of 43%.

event study will fail to find a statistically significant price impact. Notably, this error rate is many times greater than the 5% Type I error rate.

As this example illustrates, the Type II error rate that results from insisting on a Type I error rate of 5% can be quite high. Even leaving aside the question of whether a 5% significance level is consistent with applicable legal standards, we see no reason to assume that this significance level reflects the normatively appropriate trade-off.<sup>241</sup> The 5% Type I error rate is traditionally used in the academic literature on financial economics,<sup>242</sup> but there are numerous differences between those academic event studies and the ones used in securities litigation. As we have already seen, the one-sided–two-sided distinction is one such difference, as is the frequent existence of multiple relevant event dates.

In addition, most academic event studies average event date excess returns over multiple firms. This averaging often will both (i) greatly reduce the standard deviation of the statistic that is used to test for statistical significance,<sup>243</sup> and (ii) greatly reduce the importance of non-normality.<sup>244</sup> Thus, the event studies typically of interest to scholars in their academic work are atypical of event studies that are used in securities litigation. Whatever the merits of the convention of insisting on a Type I error rate of 5% in academic event studies, we think the use of that rate in securities litigation is the result of happenstance and inertia rather than either attention to legal standards or careful weighing of the costs and benefits of the trade-off in Type I and Type II errors.

This observation suggests that the current approach to using event studies in securities litigation warrants scrutiny. As long as courts continue to insist on a Type I error rate of 5%,<sup>245</sup> Type II error rates in securities litigation will be very high. This means that event study evidence of a significant price impact is much more convincing than event study evidence that fails to find a significant price impact. To put it in evidence-law terms, at the current 5% Type I error rate, a finding of significant price impact is considerably more probative than a failure to find significant price impact.

That raises two questions. First, what Type I error rate *should* courts insist on, and how should they determine that rate? Second, if event study evidence against a significant price impact has limited probative value, does

241. Fox, Fox & Gilson, *supra* note 176, at 368–72 (reaching this same conclusion).

242. See Brav & Heaton, *supra* note 11, at 599 n.31 (citing *United States v. Hatfield*, 795 F. Supp. 2d 219, 234 (E.D.N.Y. 2011), in which the court questioned whether it was appropriate to apply a 95% confidence interval when using a preponderance standard).

243. See Brav & Heaton, *supra* note 11, at 604 (“[T]he standard deviation of a sample mean’s distribution . . . falls as the number of observations reflected in the sample mean increases.”).

244. See Gelbach, Helland & Klick, *supra* note 19, at 509–10 (explaining and analyzing the standard regression approach to estimating event effects).

245. See, e.g., *In re Intuitive Surgical Sec. Litig.*, No. 5:13-cv-01920-EJD, 2016 WL 7425926, at \*15 (N.D. Cal. Dec. 22, 2016) (rejecting the conclusion of the plaintiffs’ expert based on a 90% confidence level).

that change the way courts should approach *other* evidence that is usually thought to have limited probative value? For example, one approach might be to allow financial-industry professionals to be qualified as experts for purposes of testifying that an alleged corrective disclosure could be expected to cause price impact, both for the class certification purposes on which we have focused and as to other merits questions. The logic of this idea is simple: when event study evidence fails to find a significant price impact, that evidence has limited probative value, so the value of general, nonstatistical expert opinions will be comparatively greater in such cases than in those cases in which event study evidence does find a significant price impact.<sup>246</sup> These are complex questions that go to the core of the appropriate role of event studies in securities fraud litigation and the appropriate choice of significance level.<sup>247</sup>

### Conclusion

Event studies play an important role in securities fraud litigation. In the wake of *Halliburton II*, that role will increase because proving price impact has become a virtual requirement to secure class certification. This Article has explained the event study methodology and explored a variety of considerations related to the use of event studies in securities fraud litigation, highlighting the ways in which the litigation context differs from the empirical context of many academic event studies.

A key lesson from this Article is that courts and experts should pay more attention to methodological issues. We identify four methodological considerations and demonstrate how they can be addressed. First, because a litigation-relevant event study typically involves only a single firm, issues related to non-normality of a stock's returns arise. Second, because the plaintiff must show either that the price dropped or rose but will never carry its burden if the opposite happened, experts should unquestionably be using one-sided significance testing rather than the conventionally deployed two-sided approach. Third, securities fraud litigation often involves multiple test dates, which has important and tricky implications for the appropriate level of date-specific confidence levels if the goal is an overall confidence level equal to the 95% level, which courts and experts say it is. Fourth, event studies must be modified appropriately to account for the possibility that stock price volatility varies across time.

---

246. Further, such an approach would reduce the incentive for managers to release bad news strategically in ways that would defeat the usefulness of event studies (*see supra* subpart V(B)) since doing so could open the door to more subjective expert testimony that is likely to be easy for plaintiffs to obtain.

247. A full discussion of the normative implications of the 5% Type I error rate is beyond the scope of this Article. Two of us are presently working on this question in ongoing work.



Even with these adjustments, event studies have their limits. We discuss some evidentiary challenges that confront the use of event studies in securities litigation. First, it is not clear that the 5% significance level is appropriate in litigation. Second, failing to reject the null hypothesis is not the same as proving that information did not have a price effect. As a result, the legal impact of an event study may depend critically on which party bears the burden of proof and the extent to which courts permit the introduction of non-event study evidence on price impact. Third, both accidental and strategic bundling of news may make event study evidence more difficult to muster. Fourth, event studies used in securities litigation are likely to be plagued by very high ratios of false negatives to false positives—that is, they are much more likely to yield a lack of significant evidence of an actual price impact than they are to yield significant evidence of price impact when there really was none. This imbalance of Type II and Type I error rates warrants further analysis.



*APPENDIX MATERIAL FOR*  
**AFTER HALLIBURTON: EVENT STUDIES AND  
THEIR ROLE IN FEDERAL SECURITIES FRAUD  
LITIGATION**

**August 10, 2016**

*Jill E. Fisch*<sup>\*</sup>

*Jonah B. Gelbach*<sup>\*\*</sup>

*Jonathan Klick*<sup>\*\*\*</sup>

This document contains the appendix material for “After *Halliburton*: Event Studies and Their Role in Federal Securities Fraud Litigation,” which is intended for posting as an online supplement to the main Article.

---

<sup>\*</sup> Perry Golkin Professor of Law, University of Pennsylvania Law School.

<sup>\*\*</sup> Professor of Law, University of Pennsylvania Law School.

<sup>\*\*\*</sup> Professor of Law, University of Pennsylvania Law School.

We thank Bernard Black, Ryan Bubb, Merritt Fox, Jerold Warner and participants at Rutgers-Camden, the February 2016 Penn/NYU Law and Finance Symposium, and USC-Gould for their comments, questions, and suggestions.

**APPENDIX A: IMPLEMENTING THE SQ TEST**

For convenience and comparison's sake, Table A1 summarizes the steps necessary to implement the standard approach assuming normality and the one-sided SQ test approach. As the table shows, the only difference between the two methods is in step 3, concerning how the test's critical value is calculated. The standard approach relies on the assumed normality of the estimated excess returns distribution to determine this critical value, and uses the wrong level of confidence. By contrast, the SQ test uses the data to guide the expert regarding what critical value to use.

Finally, the SQ test adds no assumptions that are not already relied on by the standard approach.<sup>1</sup> Any time the standard approach is reliable—i.e., when normality and the other assumptions addressed here are all satisfied—the SQ test is reliable as well. Thus we emphasize that there are circumstances in which the one-sided SQ test will perform systematically better than the standard approach, and there are no circumstances under which it will systematically perform worse. In sum, (i) the standard approach is generally not reliable due to the non-normality of excess returns, but (ii) the SQ test is reliable even with non-normality, and (iii) the SQ test adds virtually no additional work for an expert or court to understand. There is no good reason to use the standard approach, but not to use the SQ test. And, leaving aside the other issues we are about to discuss, there is good reason *to* use the SQ test.

---

<sup>1</sup> See note 165 at page 45 of our main Article.

**Table A1: Comparing the Standard Approach and the SQ Test Approach for Testing the Null Hypothesis of No Event Date Effect on an Alleged Corrective Disclosure Date**

<b>Step</b>	<b>Standard Approach</b>	<b><i>SQ Test Approach</i></b>
<b>1. Estimate market model</b>	Use ordinary least squares estimate of regression of Halliburton's daily stock return on daily return of market index variable(s).  (Use only data for a period that does not include any event dates in question.)	<i>Same as standard approach.</i>
<b>2. Calculate estimated excess return for event date</b>	The event date's estimated excess return equals the actual event-date return for Halliburton minus the predicted value based on the coefficients from the market model from step 1.	<i>Same as standard approach.</i>
<b>3. Calculate critical value</b>	Critical value is product of -1.96 (if two-sided critical value used) or -1.645 (if one-sided critical value used) and the standard error of regression from market model.	<i>Using estimated coefficients from the market model, calculate the estimated excess return for all dates in the estimation period from step 1. Then find the 5<sup>th</sup> percentile of these estimated excess returns. This number is the critical value.</i>
<b>4. Determine whether to reject null hypothesis of no event date effect</b>	Reject null hypothesis if event-date estimated excess return from step 2 is more negative than critical value from step 3.  Otherwise do not reject null hypothesis.	<i>Same as standard approach (except that critical value is found as described in cell just above).</i>

**APPENDIX B: CORRECTING FOR MULTIPLE TESTING WHEN THERE ARE MULTIPLE PAIRS OF ALLEGED MISREPRESENTATION AND ALLEGED CORRECTIVE DISCLOSURE DATES**

Here we briefly consider a situation like the following, in which we are concerned with proper statistical significance testing as to the merits. Plaintiff alleges that:

- on Date 1A, defendant made a non-confirmatory fraudulent statement, and then there was a corrective disclosure on Date 1B;
- on Date 2A, defendant made another non-confirmatory fraudulent statement, and then there were corrective disclosures on Dates 2B, 2C, 2D, and 2E;
- on Date 3A, defendant made a confirmatory fraudulent statement, and there was a corrective disclosure on Date 3B.

Suppose for the moment that all three alleged misrepresentations are within a common scope of claim preclusion, so that under our analysis in Part IV.C.2 of our main Article, statistical significance methodology should account for the presence of all dates related to the three alleged misrepresentations. The plaintiff will win as to alleged misrepresentation 1 if she carries her burden as to Date Set 1 (Date 1A *and* Date 1B); *or* as to Date Set 2 (Date 2A *and any one of* Dates 2B, 2C, 2D, *or* 2E); *or* as to Date set 3 (Date 3B). We assume that the court wants to ensure that there is a 5% chance of erroneously finding for the plaintiff as to any of the three alleged misrepresentations.

Following the logic of Part IV.C.3 of our main Article, the court should require a Type I error rate of 1.7% with respect to each Date Set considered separately, because if the court does that, there will be a 5% chance that at least one of the Date Sets is erroneously found to be statistically significant.<sup>2</sup> Date Set 1 involves two dates, both of which must be meet the relevant statistical threshold to for Date Set 1 to be found significant. With a Type I error rate of 1.7% for Date Set 1 overall,

---

<sup>2</sup> To be precise, the Type I error rate is 1.69524275%. To see this, observe that the probability that at least one of the 3 Date Sets will erroneously be found statistically significant is 1 minus the probability that none of them will. Using the above Type I error rate, the probability equals  $1 - (1 - 0.0169524275)^3$ , which equals 0.05, or 5%.

the appropriate Type I error rate to apply separately to Dates 1A and 1B is 13%.<sup>3</sup>

Date Set 2 involves five dates. For the plaintiff to win as to this date set, she must show that the estimated excess return meets the court's threshold as to Date 2A and any one of Dates 2B, 2C, 2D, or 2E. The court can achieve its desired overall Type I error rate for Date Set 2 by requiring a 13% Type I error rate for Date 2A considered separately, together with a 13% Type I error rate for any one of Dates 2B, 2C, 2D, or 2E. As with Date Set 1, this approach will achieve an overall Type I error rate of 1.7% for Date Set 2.<sup>4</sup> The 13% Type I error rate for any one of Dates 2B, 2C, 2D, or 2E can be achieved by requiring a Type I error rate of approximately 3.4% for each of these four dates considered separately.<sup>5</sup>

Finally, Date Set 3 involves only one date, Date 3B (since the alleged misrepresentation on Date 3A is confirmatory). Thus the appropriately  $p$ -value for Date 3B considered separately is 1.7%.

We summarize these findings for this example in Table B1. The key take-away point from this discussion is that taking appropriate account for multiple testing in securities litigation is fact-bound, depending importantly on how the various alleged events interact with each other according to the plaintiff's allegations.<sup>6</sup>

---

<sup>3</sup> Since the probability that both Date 1A and Date 1B will have an estimated excess return with a  $p$ -value of  $q$  is  $q^2$ , and since we want  $q$  to equal 0.017, we have  $q = \sqrt{0.017}$ , which is approximately 0.13, or 13%.

<sup>4</sup> The math is the same as in footnote 3 of the main Article; the only difference is that the 13% Type I error rate for Date 1B is replaced by a Type I error rate of 13% for finding any of Dates 2B, 2C, 2D, or 2E significant.

<sup>5</sup> This is so because the probability that at least one of the 4 Dates Date 2A, 2B, 2C, or 2D will erroneously be found statistically significant is 1 minus the probability that none of them will. Using the above Type I error rate, the probability equals  $1 - (1 - 0.03421642936)^4$ , which equals roughly 0.034, or 3.4%.

<sup>6</sup> Careful readers will have noted that the discussion in this section has not made use of Holm's sequential approach to determining which Date Sets, and which of the Dates within them, should be treated as statistically significant. The discussion in this section is thus the simple, non-sequential Šidák approach. Using Holm's sequential approach would involve greater power and should be done in practice; we have not used it in this section since it takes more notation to explain and illustrate.

**Table B1: Separately Considered  $p$ -Values that Yield a 5% Type I Error Rate in Finding for the Plaintiff as to Any Alleged Misrepresentation Date of the Three Date Sets Considered in Our Example**

Type I Error Rate for Each Date Considered Separately	
<i>Date Set 1</i>	
<b>Date 1A</b>	<b>13%</b>
<b>Date 1B</b>	<b>13%</b>
(Both dates' $p$ -values must be less than 0.13 for plaintiff to win as to alleged misrepresentation on Date 1A.)	
<i>Date Set 2</i>	
<b>Date 2A</b>	<b>13%</b>
<b>Date 2B</b>	<b>3.4%</b>
<b>Date 2C</b>	<b>3.4%</b>
<b>Date 2D</b>	<b>3.4%</b>
<b>Date 2E</b>	<b>3.4%</b>
(Date 2A $p$ -value must be less than 0.13 and $p$ -value must be less than 0.034 for at least one of Dates 2B-2E for plaintiff to win as to alleged misrepresentation on Date 2A.)	
<i>Date Set 3</i>	
<b>Date 3A</b>	Confirmatory alleged misrepresentation, so no relevant data
<b>Date 3B</b>	<b>1.7%</b>
(Date 3B $p$ -value must be less than 0.017 for plaintiff to win as to alleged misrepresentation on Date 3A.)	

### APPENDIX C: ALLOWING FOR HETEROGENEITY IN THE DISTRIBUTION OF EXCESS RETURNS

In this appendix we drop the assumption that excess returns have the same distribution on every date. In doing so, we allow for two possible reasons why the variance of daily excess returns might vary across dates. The first is that variance of daily excess returns might be associated with contemporaneous and lagged variance in the daily returns of the industry index variables we use in the market model described in Part III of our main Article, as well as an index of overall volatility in the stock market.<sup>7</sup> The second source of heterogeneity follows a long literature that models the variance of securities using a model known as generalized autoregressive conditional heteroskedasticity (“GARCH”).<sup>8</sup> The GARCH model class allows the variance of a security’s return on a given date to be related both to the variance on earlier dates (this is the autoregressive part) and to the magnitude of earlier dates’ realized values of squared excess returns (this is the conditional heteroskedasticity part).

#### A. A Model that Allows for Heterogeneity in the Variance of Excess Returns

In the absence of any special news on date  $t$ , our basic model may be written as

$$H_t = \alpha + M_t' \beta + \varepsilon_t,$$

where  $t$  is the date;  $H_t$  is Halliburton’s daily stock return;  $\alpha$  is the intercept in the market model;  $M_t$  is a (column) vector of market index returns (such as the three included in our market model, discussed in Part

---

<sup>7</sup> This index is the VIX index published by the Chicago Board Options Exchange. This index, sometimes known as the “fear index”, may be associated with the volatility of firms’ daily returns, which can be accounted for by modeling individual stock variance to take account of movements in the VIX. See Andrew C. Baker, Note, *Single-Firm Event Studies, Securities Fraud, and Financial Crisis: Problems of Inference*, 68 Stan. L. Rev. 1207, 1245 (2016). We include this index here to account for such concerns.

<sup>8</sup> The GARCH model was first developed in detail in Timothy Bollerslev, *Generalized Autoregressive Conditional Heteroskedasticity*, 31 J. Econometrics 307 (1986), building on seminal work by Robert Engle, *Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of UK Inflation*, 50 Econometrica 987 (1982). For an excellent and generally accessible discussion, see Chapter 12 of John Y. Campbell, Andrew W. Lo, and A. Craig MacKinlay, *THE ECONOMETRICS OF FINANCIAL MARKETS* (1997). For a use of GARCH in the law and finance literature, see Mark I. Weinstein, *Don’t Buy Shares without It: Limited Liability Comes to American Express*, 37 J. Leg. Stud. 189 (2008).



IV.D of our main Article);  $\beta$  is a vector of coefficients that link these indexes' returns to Halliburton's return; and  $\varepsilon_t$  is the excess return for date  $t$  given that no special news occurred.

To accommodate these possible sources of variance heterogeneity, we write the daily excess return  $\varepsilon_t$  as the product of two components:

$$\varepsilon_t = \sigma_t u_t,$$

where  $u_t$  is a random variable that has mean 0 and variance 1. This variable might or might not be normally distributed, but by assumption it is independently and identically distributed ("iid"); importantly, the independence assumed here is universal, in the sense that we assume  $u_t$  has no association with any other variable in the model—it is pure "white noise".

Let  $I_t$  denote the set of all information available on date  $t$ . The variable  $\sigma_t$  is known as the conditional variance of the daily excess return, because the variance of the daily excess return on date  $t$ —given all information available as of that date—is  $E[\varepsilon_t^2 | I_t] = E[(\sigma_t u_t)^2 | I_t]$ ,<sup>9</sup> which can be shown to equal  $E[\sigma_t^2 | I_t]$ .<sup>10</sup> The approach taken until Part IV.D of our main Article can be regarded as being equivalent to assuming that the variance does not actually change across dates, so that  $\sigma_t = \sigma$  for all dates  $t$ . In that case we would have  $\varepsilon_t = \sigma u_t$ , with  $u_t$  simply being the result of defining the standardized daily excess return, i.e.,  $u_t \equiv \varepsilon_t / \sigma_t$ .

We allow for the possibility that  $\sigma_t$  varies over time through three channels. First, we allow for the possibility that the date- $t$  conditional variance of the date- $t$  excess return,  $\sigma_t^2$ , varies with the value of the squared returns for our three industry index variables.<sup>11</sup> Second, we

<sup>9</sup> The variance of  $\varepsilon_t$  equals the expected value of its square since  $\varepsilon_t$  has mean zero.

<sup>10</sup> This follows because  $\text{Variance}[\varepsilon_t] = \text{Variance}[\sigma_t u_t] = E[(\sigma_t u_t)^2] - (E[\sigma_t u_t])^2$ . Since  $u_t$  is independent of everything and has mean zero, we have  $E[\sigma_t u_t] = 0$ . Further,  $E[(\sigma_t u_t)^2] = E_{u_t}\{E[\sigma_t^2 | u_t] u_t^2\} = E[\sigma_t^2] E[u_t^2] = E[\sigma_t^2]$ . The first equality follows by the law of iterated expectations; the second follows because  $\sigma_t$  and  $u_t$  are independent by assumption; and the third follows because  $u_t$  has variance 1.

<sup>11</sup> Contemporaneous values of the squared returns of the indexes—values for date  $t$ —are not yet known during date  $t$ . Thus if we were trying to estimate a model of daily excess returns from the perspective of investors, it would be inappropriate to include the contemporaneous squared index returns in our model explaining  $\sigma_t$ . However, in conducting an event study used for securities fraud litigation, it is appropriate to take an ex post view by including these variables: we are attempting to determine whether event-date excess returns are unusual enough to be considered statistically significant, given not only the information available to investors when they traded on event dates,

allow for the possibility that  $\sigma_t^2$  varies with the value of the VIX volatility index.<sup>12</sup> And third, we allow for the possibility that  $\sigma_t^2$  varies with the GARCH terms described above, namely the one-period lags  $\sigma_{t-1}^2$  and  $\varepsilon_{t-1}^2$ . Our model for  $\sigma_t^2$  is:

$$\sigma_t^2 = \sigma_0^2 + V_t' \theta_1 + \sigma_{t-1}^2 \theta_2 + \varepsilon_{t-1}^2 \theta_3,$$

where  $V_t$  is a column vector of the squared returns for the industry peer index, our version of the S&P 500 energy sector index, our version of the energy and construction index, and the VIX index; and the various  $\theta$  parameters connect each explanatory variable on the right hand side to the daily conditional variance. Notice that if all explanatory variables were 0, we would have constant conditional variance with  $\sigma_t^2 = \sigma_0^2$ , which is why we give the intercept in the conditional variance equation the name  $\sigma_0^2$ . Finally, we note that since the standard deviation is the square of the variance, the above model for the daily variance  $\sigma_t^2$  implicitly is also a model for the daily standard deviation  $\sigma$ .

With this model in hand, the next question is how to estimate it. The model is obviously interconnected with our market model, since both involve the daily excess return  $\varepsilon_t$ .<sup>13</sup> If we were willing to assume not just that the  $u_t$  component of the daily excess return is iid, but also that it follows a normal distribution, then it would be straightforward to write down the model's log likelihood function and maximize it with respect to the parameters  $\alpha, \beta, \sigma_0^2, \theta_1, \theta_2$ , and  $\theta_3$ ; many statistical software packages such as Stata do this using canned routines that are easy to invoke.<sup>14</sup> We estimate the model under this assumption and then return to the question of non-normality below.

---

*but also available information about what else happened that date for related securities.*

<sup>12</sup> This index is the VIX index published by the Chicago Board Options Exchange. This index, sometimes known as the “fear index”, may be associated with the volatility of firms’ daily returns, which can be accounted for by modeling individual stock variance to take account of movements in the VIX.

<sup>13</sup> We can write the market model as

$$H_t = \alpha + M_t \beta + \varepsilon_t,$$

where  $M_t$  is a vector of the industry index variables and  $\alpha$  and  $\beta$  are parameters.

<sup>14</sup> In Stata, the routine in question is the “arch” command, using the options “arch(1)”, “garch(1)”, and “het(...)”, where the ellipsis in the last option is replaced with a list of names of the variables in  $V_{2t}$  from our equations in the text.

*B. Estimation Results*

Here we report the results from estimating the model described in section A of this appendix. Table C1 reports the coefficient estimates for the parameters in the market model, i.e., the betas that explain Halliburton's daily return. The coefficient estimates and estimated standard errors are very similar to those in Table 2 of our main Article, as should be expected given the robustness result described in footnote 19, *infra*.

**Table C1: Estimates from Market Model for Excess Return, from Model that Allows Heterogeneity in Conditional Variance**

	Market Model for Excess Return	
	Coefficient	Standard error*
<i><math>\beta</math> parameters</i>		
Industry Peers	0.941	0.034
S&P 500 Energy	0.162	0.055
Engineering & Construction	0.026	0.035
<i><math>\alpha</math> parameters</i>		
Intercept	-0.001	0.001
* Estimated standard errors computed using robust covariance estimator to account for possibility of non-normality in standardized excess returns, $u_t$ .		

The coefficient estimates for the conditional variance equation—the estimates of the  $\theta$  parameters, using our notation from above—appear in Table C2. The coefficient estimates are statistically insignificant for the squared daily returns of the S&P 500 Energy index and the Engineering & Construction Index (further testing indicated that they are also jointly insignificant). However, coefficients are highly statistically significant for both the VIX index and the industry peer index squared daily returns, which tells us that there is market-associated variation in the conditional variance of Halliburton's excess return. The coefficient on the squared returns of the industry peer index (0.330) is several times the magnitude of the VIX coefficient (0.080), indicating that volatility in the excess returns of Halliburton's direct competitors plays a comparatively more important role than the volatility in the stock market overall that VIX

measures.<sup>15</sup> The coefficients on the GARCH parameters show that there is a statistically significant coefficient on the autoregressive term (i.e., the lagged variance term). The conditional heteroscedasticity coefficient (i.e., the coefficient on the square of lagged excess returns) is twice the size of this coefficient, though it is statistically significant only at the 10% level.

**Table C2: Estimates from Market Model for Conditional Variance**

	Coefficient	Standard error*
<i>Square of daily return for:</i>		
Industry Peers**	0.330	0.083
S&P 500 Energy**	0.068	0.080
Engineering & Construction**	0.080	0.061
<i>Level of:</i>		
VIX**	0.044	0.017
<i>GARCH lag parameters for:</i>		
Autoregressive term (lagged variance)	0.128	0.062
Conditional heteroskedasticity term (square of lag in realized excess return)	0.244	0.147
<i>Intercept:</i>		
$\sigma_0^2$	-10.024	0.520
* Estimated standard errors computed using robust covariance estimator to account for possibility of non-normality in standardized excess returns, $u_t$ .		
** Variables are scaled so that all these coefficients represent the effect of an increase of one-unit of standard deviation of each regressor.		

<sup>15</sup> Because the VIX index and all squared returns variables have been divided by their respective sample standard deviations, the coefficient magnitudes may meaningfully be compared. Thus, the effect of a one-unit movement in the VIX variable included in our model is comparable to the effect of a one-unit movement in the square of each of our industry indexes.

We used the coefficient estimates from Table C1 to estimate daily excess returns for both those dates in the estimation period and those dates on which alleged corrective disclosures occurred. That is, we used these coefficients to estimate the set of  $\varepsilon_t$  values:

$$\hat{\varepsilon}_t = H_t - \hat{\alpha} - M_t \hat{\beta},$$

where hats denote estimates and we recall that  $H_t$  and  $M_t$  are, respectively, the observed date- $t$  returns for Halliburton and the observed vector of date- $t$  returns for the market indexes.

To estimate the conditional daily variance  $\sigma_t^2$ , we used

$$\hat{\sigma}_t^2 = \hat{\sigma}_0^2 + V_t' \hat{\theta}_1 + \hat{\sigma}_{t-1}^2 \hat{\theta}_2 + \hat{\varepsilon}_{t-1}^2 \hat{\theta}_3,$$

where hats again denote estimates.<sup>16</sup> (The estimated conditional daily standard deviation  $\hat{\sigma}_t$  may be calculated by taking the square root of the estimated conditional daily variance.)

### C. Statistical Significance Testing Methodology

In this section we discuss how to use the estimated standardized estimated excess returns, computed in section B just above, to test for statistical significance of the event date estimated excess return. We first construct the estimated scale factor discussed in Part IV.D of our main Article. To do so for some date  $t$ , we first calculate the conditional daily variance for the event date,  $\hat{\sigma}_*^2$  and the conditional daily variance for date  $t$ ,  $\hat{\sigma}_t^2$ . The estimated scale factor is the square-root of the ratio of these estimates:

$$\widehat{SF}_t \equiv \sqrt{\frac{\hat{\sigma}_*^2}{\hat{\sigma}_t^2}}.$$

---

<sup>16</sup> The only trick here is that we need estimates of  $\hat{\sigma}_t^2$  and  $\hat{\varepsilon}_t^2$  for the date before the initial date in our estimation sample. We use Stata's default approach of using the estimated unconditional variance; see Stata Corporation, *Autoregressive conditional heteroskedasticity (ARCH) family of estimators*, <http://www.stata.com/manuals13/tsarch.pdf> at 12-13 ("Priming" tab, "arch0" option defaults). This is the method suggested in Timothy Bollerslev, *Generalized Autoregressive Conditional Heteroskedasticity*, 31 J. Econometrics 307 (1986); see also Matteo M. Pelagatti and Francesco Lisis, *Variance Initialisation in GARCH Estimation*, paper presented at Complex data modeling and computationally intensive statistical methods for estimation and prediction, Milan September 14-16, 2009, available at <http://www2.mate.polimi.it/ocs/viewpaper.php?id=127&cf=7>.

We then create the re-scaled estimated excess return as

$$\hat{r}_t \equiv \widehat{SF}_t \times \hat{\varepsilon}_t,$$

where  $\hat{\varepsilon}_t$  is the estimated excess return for date  $t$ . Under the null hypothesis that there was no special price impact on the event date, the re-scaled estimated excess return  $\hat{r}_t$  has the same limiting distribution as the event date estimated excess return. Consequently, we may use the collection of  $\hat{r}_t$  estimates for testing statistical significance of the event date estimated excess return just as we would use the unscaled estimated excess returns in the absence of variance heterogeneity. Which tests are appropriate depend on the distribution of the underlying white noise term  $u_t$ .

It is helpful to first consider what one might do if the white noise term  $u_t$  were normal. In this case, the re-scaled excess returns—i.e., the  $r_t$ —are themselves normal; all non-normality in the unconditional distribution of  $\varepsilon_t$  is due to the heterogeneity in the daily standard deviation  $\sigma_t$ .<sup>17</sup> The usual critical values based on the normal distribution are justified when the normality assumption holds.<sup>18</sup> However, this assumption could be wrong, and as in Part IV.B of our main Article for the homogeneous variance case, a method of testing for statistical significance that ignores actually existing non-normality is unreliable.

Fortunately, there is a simple and robust alternative, due to a famous result in the econometric theory literature. Even when  $u_t$  is not actually distributed normally, under very broad conditions it turns out that the model parameters  $\alpha, \beta, \sigma_0^2, \theta_1, \theta_2$ , and  $\theta_3$  will be appropriately estimated if we use the maximum likelihood estimator that would follow from incorrectly assuming normality of  $u_t$ .<sup>19</sup>

For any date  $t$ , consider the estimated white noise term  $\hat{u}_t \equiv \frac{\hat{\varepsilon}_t}{\hat{\sigma}_t}$ . Since  $\hat{\varepsilon}_t$  and  $\hat{\sigma}_t^2$  are appropriate estimates of  $\varepsilon_t$  and  $\sigma_t^2$ , respectively, it follows

<sup>17</sup> See, e.g., discussion in Chapter 12 of John Y. Campbell, Andrew W. Lo, and A. Craig MacKinlay, *THE ECONOMETRICS OF FINANCIAL MARKETS* 480-81 (1997) (explaining that even when the variable we have called  $u_t$  is normal, heterogeneity in  $\sigma_t$  will necessarily cause excess kurtosis, and thus non-normality, in  $\varepsilon_t$ ).

<sup>18</sup> This is the approach taken by Baker, note 7, *supra*.

<sup>19</sup> See Tim Bollerslev and Jeffrey M. Wooldridge, *Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances*, 11 *Econometric Reviews* 143 (1992) (proving that the so-called Quasi-Maximum Likelihood Estimator that results “when a normal log-likelihood is maximized but the assumption of normality is violated” yields a consistent and asymptotically normal estimator of parameter estimates under broad conditions and specifically analyzing GARCH models).

that  $\hat{u}_t$  is an appropriate estimate of the true white noise terms  $u_t$  defined above.<sup>20</sup> This means that we may use the empirical distribution of estimated white noise terms to estimate features of the distribution of the true white noise term. We then used Stata's **sktest** command to test for normality of the white noise term. The test rejected normality resoundingly, with a  $p$ -value of 0.025, due to substantial kurtosis.<sup>21</sup>

The “usual”  $p$ -values computed in Part IV.B of our main Article were then computed using the SQ test applied to the event date estimated excess return, using the sample quantiles of the re-scaled estimated excess returns. That is, each event date's “usual”  $p$ -value equals the share of estimation-period dates that had re-scaled estimated excess returns greater than the event date estimated excess return.

---

<sup>20</sup> Formally, the Slutsky theorem tells us that the probability limit passes through a continuous function, so that a continuous function of consistent estimators is consistent for the continuous function of the probability limits of those estimators. Here that means that  $\hat{u}_t$  is consistent for  $u_t$ .

<sup>21</sup> We note in addition that the 5<sup>th</sup> sample quantile of the estimated white noise term (-1.742) was somewhat more negative than the 5<sup>th</sup> quantile of the standard normal distribution (-1.645), which is relevant to the comparative performance of the SQ test and the usual test assuming normality of the white noise term. As discussed in section IV.B of our main Article, the empirical magnitude of such non-normality will vary with context.



# Exhibit 53

# Reference Manual on Scientific Evidence

*Third Edition*

Committee on the Development of the Third Edition of the  
Reference Manual on Scientific Evidence

Committee on Science, Technology, and Law  
Policy and Global Affairs

FEDERAL JUDICIAL CENTER

NATIONAL RESEARCH COUNCIL  
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS  
Washington, D.C.  
**[www.nap.edu](http://www.nap.edu)**

**THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001**

The Federal Judicial Center contributed to this publication in furtherance of the Center's statutory mission to develop and conduct educational programs for judicial branch employees. The views expressed are those of the authors and not necessarily those of the Federal Judicial Center.

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

The development of the third edition of the *Reference Manual on Scientific Evidence* was supported by Contract No. B5727.R02 between the National Academy of Sciences and the Carnegie Corporation of New York and a grant from the Starr Foundation. The views expressed in this publication are those of the authors and do not necessarily reflect those of the National Academies or the organizations that provided support for the project.

International Standard Book Number-13: 978-0-309-21421-6

International Standard Book Number-10: 0-309-21421-1

Library of Congress Cataloging-in-Publication Data

Reference manual on scientific evidence. — 3rd ed.

p. cm.

Includes bibliographical references and index.

ISBN-13: 978-0-309-21421-6 (pbk.)

ISBN-10: 0-309-21421-1 (pbk.)

1. Evidence, Expert—United States. I. Federal Judicial Center.

KF8961.R44 2011

347.73'67—dc23

2011031458

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, <http://www.nap.edu>.

Copyright 2011 by the National Academy of Sciences. All rights reserved.

Printed in the United States of America

# THE FEDERAL JUDICIAL CENTER

The Federal Judicial Center is the research and education agency of the federal judicial system. It was established by Congress in 1967 (28 U.S.C. §§ 620–629), on the recommendation of the Judicial Conference of the United States, with the mission to “further the development and adoption of improved judicial administration in the courts of the United States.” By statute, the Chief Justice of the United States chairs the Federal Judicial Center’s Board, which also includes the director of the Administrative Office of the U.S. Courts and seven judges elected by the Judicial Conference.

The Center undertakes empirical and exploratory research on federal judicial processes, court management, and sentencing and its consequences, often at the request of the Judicial Conference and its committees, the courts themselves, or other groups in the federal system. In addition to orientation and continuing education programs for judges and court staff on law and case management, the Center produces publications, videos, and online resources. The Center provides leadership and management education for judges and court employees, and other training as needed. Center research informs many of its educational efforts. The Center also produces resources and materials on the history of the federal courts, and it develops resources to assist in fostering effective judicial administration in other countries.

Since its founding, the Center has had nine directors. Judge Barbara J. Rothstein became director of the Federal Judicial Center in 2003

**[www.fjc.gov](http://www.fjc.gov)**

# **THE NATIONAL ACADEMIES**

## ***Advisers to the Nation on Science, Engineering, and Medicine***

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

**[www.national-academies.org](http://www.national-academies.org)**

*Committee on the Development of the Third Edition of the  
Reference Manual on Scientific Evidence*

*Co-Chairs:*

**JEROME P. KASSIRER** (IOM), Distinguished Professor, Tufts University  
School of Medicine

**GLADYS KESSLER**, Judge, U.S. District Court for the District of Columbia

*Members:*

**MING W. CHIN**, Associate Justice, The Supreme Court of California

**PAULINE NEWMAN**, Judge, U.S. Court of Appeals for the Federal Circuit

**KATHLEEN MCDONALD O'MALLEY**, Judge, U.S. Court of Appeals for  
the Federal Circuit

**JED S. RAKOFF**, Judge, U.S. District Court, Southern District of New York

**CHANNING R. ROBERTSON**, Ruth G. and William K. Bowes Professor,  
School of Engineering, and Professor, Department of Chemical Engineering,  
Stanford University

**JOSEPH V. RODRICKS**, Principal, Environ

**ALLEN WILCOX**, Senior Investigator, Institute of Environmental Health  
Sciences

**SANDY L. ZABELL**, Professor of Statistics and Mathematics, Weinberg  
College of Arts and Sciences, Northwestern University

*Consultant to the Committee:*

**JOE S. CECIL**, Project Director, Program on Scientific and Technical Evidence,  
Division of Research, Federal Judicial Center

*Staff:*

**ANNE-MARIE MAZZA**, Director

**STEVEN KENDALL**, Associate Program Officer

**GURUPRASAD MADHAVAN**, Program Officer (until November 2010)

*Board of the Federal Judicial Center*

The Chief Justice of the United States, *Chair*

Judge Susan H. Black, U.S. Court of Appeals for the Eleventh Circuit

Magistrate Judge John Michael Facciola, U.S. District Court for the District of  
Columbia

Judge James B. Haines, U.S. Bankruptcy Court for the District of Maine

Chief Judge James F. Holderman, U.S. District Court for the Northern District  
of Illinois

Judge Edward C. Prado, U.S. Court of Appeals for the Fifth Circuit

Chief Judge Loretta A. Preska, U.S. District Court for the Southern District of  
New York

Chief Judge Kathryn H. Vratil, U.S. District Court for the District of Kansas

James C. Duff, Director of the Administrative Office of the U.S. Courts



*Committee on Science, Technology, and Law*  
*National Research Council*

**DAVID KORN** (*Co-Chair*), Professor of Pathology, Harvard Medical School, and formerly, Inaugural Vice Provost for Research, Harvard University

**RICHARD A. MESERVE** (*Co-Chair*), President, Carnegie Institution for Science, and Senior of Counsel, Covington & Burling LLP

**FREDERICK R. ANDERSON, JR.**, Partner, McKenna, Long & Aldridge LLP

**ARTHUR I. BIENENSTOCK**, Special Assistant to the President for Federal Research Policy, and Director, Wallenberg Research Link, Stanford University

**BARBARA E. BIERER**, Professor of Medicine, Harvard Medical School, and Senior Vice President, Research, Brigham and Women's Hospital

**ELIZABETH H. BLACKBURN**, Morris Herzstein Professor of Biology and Physiology, University of California, San Francisco

**JOHN BURRIS**, President, Burroughs Wellcome Fund

**ARTURO CASADEVALL**, Leo and Julia Forchheimer Professor of Microbiology and Immunology; Chair, Department of Biology and Immunology; and Professor of Medicine, Albert Einstein College of Medicine

**JOE S. CECIL**, Project Director, Program on Scientific and Technical Evidence, Division of Research, Federal Judicial Center

**ROCHELLE COOPER DREYFUSS**, Pauline Newman Professor of Law and Director, Engelberg Center on Innovation Law and Policy, New York University School of Law

**DREW ENDY**, Assistant Professor, Bioengineering, Stanford University, and President, The BioBricks Foundation

**PAUL G. FALKOWSKI**, Board of Governors Professor in Geological and Marine Science, Department of Earth and Planetary Science, Rutgers, The State University of New Jersey

**MARCUS FELDMAN**, Burnet C. and Mildred Wohlford Professor of Biological Sciences, Stanford University

**ALICE P. GAST**, President, Lehigh University

**JASON GRUMET**, President, Bipartisan Policy Center

**BENJAMIN W. HEINEMAN, JR.**, Senior Fellow, Harvard Law School and Harvard Kennedy School of Government

**D. BROCK HORNBY**, U.S. District Judge for the District of Maine

**ALAN B. MORRISON**, Lerner Family Associate Dean for Public Interest and Public Service, George Washington University Law School

**PRABHU PINGALI**, Deputy Director of Agricultural Development, Global Development Program, Bill and Melinda Gates Foundation

**HARRIET RABB**, Vice President and General Counsel, Rockefeller  
University

**BARBARA JACOBS ROTHSTEIN**, Director, The Federal Judicial Center

**DAVID S.TATEL**, Judge, U.S. Court of Appeals for the District of Columbia  
Circuit

**SOPHIE VANDEBROEK**, Chief Technology Officer and President, Xerox  
Innovation Group, Xerox Corporation

*Staff*

**ANNE-MARIE MAZZA**, Director

**STEVEN KENDALL**, Associate Program Officer

## Foreword

In 1993, in the case *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, the Supreme Court instructed trial judges to serve as “gatekeepers” in determining whether the opinion of a proffered expert is based on scientific reasoning and methodology. Since *Daubert*, scientific and technical information has become increasingly important in all types of decisionmaking, including litigation. As a result, the science and legal communities have searched for expanding opportunities for collaboration.

Our two institutions have been at the forefront of trying to improve the use of science by judges and attorneys. In *Daubert*, the Supreme Court cited an *amicus curiae* brief submitted by the National Academy of Sciences and the American Association for the Advancement of Science to support the view of science as “a process for proposing and refining theoretical explanations about the world that are subject to further testing and refinement.” Similarly, in *Kumho Tire Co. v. Carmichael* (1999) the Court cited an *amicus* brief filed by the National Academy of Engineering for its assistance in explaining the process of engineering.

Soon after the *Daubert* decision the Federal Judicial Center published the first edition of the *Reference Manual on Scientific Evidence*, which has become the leading reference source for federal judges for difficult issues involving scientific testimony. The Center also undertook a series of research studies and judicial education programs intended to strengthen the use of science in courts.

More recently the National Research Council through its Committee on Science, Technology, and Law has worked closely with the Federal Judicial Center to organize discussions, workshops, and studies that would bring the two communities together to explore the nature of science and engineering, and the processes by which science and technical information informs legal issues. It is in that spirit that our organizations joined together to develop the third edition of the *Reference Manual on Scientific Evidence*. This third edition, which was supported by grants from the Carnegie Foundation and the Starr Foundation, builds on the foundation of the first two editions, published by the Center. This edition was overseen by a National Research Council committee composed of judges and scientists and engineers who share a common vision that together scientists and engineers and members of the judiciary can play an important role in informing judges about the nature and work of the scientific enterprise.

Our organizations benefit from the contributions of volunteers who give their time and energy to our efforts. During the course of this project, two of the chapter authors passed away: Margaret Berger and David Friedman. Both Margaret and David served on NRC committees and were frequent contributors to Center judicial education seminars. Both were involved in the development of the *Reference Manual* from the beginning, both have aided each of our institutions through their services on committees, and both have made substantial contributions to our understanding of law and science through their individual scholarship.

*Reference Manual on Scientific Evidence*

They will be missed but their work will live on in the thoughtful scholarship they have left behind.

We extend our sincere appreciation to Dr. Jerome Kassirer and Judge Gladys Kessler and all the members of the committee who gave so generously to make this edition possible.

THE HONORABLE BARBARA J. ROTHSTEIN  
*Director*  
*Federal Judicial Center*

RALPH J. CICERONE  
*President*  
*National Academy of Sciences*

## Acknowledgments

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Academies' Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, accuracy, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the process.

We wish to thank the following individuals for their review of selected chapters of this report: Bert Black, Mansfield, Tanick & Cohen; Richard Bjur, University of Nevada; Michael Brick, Westat; Edward Cheng, Vanderbilt University; Joel Cohen, Rockefeller University; Morton Corn, Morton Corn and Associates; Carl Cranor, University of California, Riverside; Randall Davis, Massachusetts Institute of Technology; John Doull, University of Kansas; Barry Fisher, Los Angeles County Sheriff's Department; Edward Foster, University of Minnesota; David Goldston, Natural Resources Defense Council; James Greiner, Harvard University; Susan Haack, University of Miami; David Hillis, University of Texas; Karen Kafadar, Indiana University; Graham Kalton, Westat; Randy Katz, University of California, Berkeley; Alan Leshner, American Association for the Advancement of Science; Laura Liptai, Biomedical Forensics; Patrick Malone, Patrick Malone & Associates; Geoffrey Mearns, Cleveland State University; John Monahan, The University of Virginia; William Nordhaus, Yale University; Fernando Olguin, U.S. District Court for the Central District of California; Jonathan Samet, University of Southern California; Nora Cate Schaeffer, University of Wisconsin; Shira Scheindlin, U.S. District Court for the Southern District of New York; and Reggie Walton, U.S. District Court for the District of Columbia.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the report, nor did they see the final draft of the report before its release. The review of this report was overseen by D. Brock Hornby, U.S. District Judge for the District of Maine. Appointed by the National Academies, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring committee and the institution.



## Preface

Supreme Court decisions during the last decade of the twentieth century mandated that federal courts examine the scientific basis of expert testimony to ensure that it meets the same rigorous standard employed by scientific researchers and practitioners outside the courtroom. Needless to say, this requirement places a demand on judges not only to comprehend the complexities of modern science but to adjudicate between parties' differing interpretations of scientific evidence. Science, meanwhile, advances. Methods change, new fields are born, new tests are introduced, the lexicon expands, and fresh approaches to the interpretation of causal relations evolve. Familiar terms such as enzymes and molecules are replaced by microarray expression and nanotubes; single-author research studies have now become multi-institutional, multi-author, international collaborative efforts.

No field illustrates the evolution of science better than forensics. The evidence provided by DNA technology was so far superior to other widely accepted methods and called into question so many earlier convictions that the scientific community had to reexamine many of its time-worn forensic science practices. Although flaws of some types of forensic science evidence, such as bite and footprint analysis, lineup identification, and bullet matching were recognized, even the most revered form of forensic science—fingerprint identification—was found to be fallible. Notably, even the “gold standard” of forensic evidence, namely DNA analysis, can lead to an erroneous conviction if the sample is contaminated, if specimens are improperly identified, or if appropriate laboratory protocols and practices are not followed.

Yet despite its advances, science has remained fundamentally the same. In its ideal expression, it examines the nature of nature in a rigorous, disciplined manner in, whenever possible, controlled environments. It still is based on principles of hypothesis generation, scrupulous study design, meticulous data collection, and objective interpretation of experimental results. As in other human endeavors, however, this ideal is not always met. Feverish competition between researchers and their parent institutions, fervent publicity seeking, and the potential for dazzling financial rewards can impair scientific objectivity. In recent years we have experienced serious problems that range from the introduction of subtle bias in the design and interpretation of experiments to overt fraudulent studies. In this welter of modern science, ambitious scientists, self-designated experts, billion-dollar corporate entities, and aggressive claimants, judges must weigh evidence, judge, and decide.

As with previous editions of the *Reference Manual*, this edition is organized according to many of the important scientific and technological disciplines likely to be encountered by federal (or state) judges. We wish to highlight here two critical issues germane to the interpretation of all scientific evidence, namely issues of causation and conflict of interest. Causation is the task of attributing cause and effect, a normal everyday cognitive function that ordinarily takes little or



no effort. Fundamentally, the task is an inferential process of weighing evidence and using judgment to conclude whether or not an effect is the result of some stimulus. Judgment is required even when using sophisticated statistical methods. Such methods can provide powerful evidence of associations between variables, but they cannot prove that a causal relationship exists. Theories of causation (evolution, for example) lose their designation as theories only if the scientific community has rejected alternative theories and accepted the causal relationship as fact. Elements that are often considered in helping to establish a causal relationship include predisposing factors, proximity of a stimulus to its putative outcome, the strength of the stimulus, and the strength of the events in a causal chain. Unfortunately, judges may be in a less favorable position than scientists to make causal assessments. Scientists may delay their decision while they or others gather more data. Judges, on the other hand, must rule on causation based on existing information. Concepts of causation familiar to scientists (no matter what stripe) may not resonate with judges who are asked to rule on general causation (i.e., is a particular stimulus known to produce a particular reaction) or specific causation (i.e., did a particular stimulus cause a particular consequence in a specific instance). In the final analysis, a judge does not have the option of suspending judgment until more information is available, but must decide after considering the best available science. Finally, given the enormous amount of evidence to be interpreted, expert scientists from different (or even the same) disciplines may not agree on which data are the most relevant, which are the most reliable, and what conclusions about causation are appropriate to be derived.

Like causation, conflict of interest is an issue that cuts across most, if not all, scientific disciplines and could have been included in each chapter of the *Reference Manual*. Conflict of interest manifests as bias, and given the high stakes and adversarial nature of many courtroom proceedings, bias can have a major influence on evidence, testimony, and decisionmaking. Conflicts of interest take many forms and can be based on religious, social, political, or other personal convictions. The biases that these convictions can induce may range from serious to extreme, but these intrinsic influences and the biases they can induce are difficult to identify. Even individuals with such prejudices may not appreciate that they have them, nor may they realize that their interpretations of scientific issues may be biased by them. Because of these limitations, we consider here only financial conflicts of interest; such conflicts are discoverable. Nonetheless, even though financial conflicts can be identified, having such a conflict, even one involving huge sums of money, does not necessarily mean that a given individual will be biased. Having a financial relationship with a commercial entity produces a conflict of interest, but it does not inevitably evoke bias. In science, financial conflict of interest is often accompanied by disclosure of the relationship, leaving to the public the decision whether the interpretation might be tainted. Needless to say, such an assessment may be difficult. The problem is compounded in scientific publications by obscure ways in which the conflicts are reported and by a lack of disclosure of dollar amounts.

*Preface*

Judges and juries, however, must consider financial conflicts of interest when assessing scientific testimony. The threshold for pursuing the possibility of bias must be low. In some instances, judges have been frustrated in identifying expert witnesses who are free of conflict of interest because entire fields of science seem to be co-opted by payments from industry. Judges must also be aware that the research methods of studies funded specifically for purposes of litigation could favor one of the parties. Though awareness of such financial conflicts in itself is not necessarily predictive of bias, such information should be sought and evaluated as part of the deliberations.

*The Reference Manual on Scientific Evidence*, here in its third edition, is formulated to provide the tools for judges to manage cases involving complex scientific and technical evidence. It describes basic principles of major scientific fields from which legal evidence is typically derived and provides examples of cases in which such evidence was used. Authors of the chapters were asked to provide an overview of principles and methods of the science and provide relevant citations. We expect that few judges will read the entire manual; most will use the volume in response to a need when a particular case arises involving a technical or scientific issue. To help in this endeavor, the *Reference Manual* contains completely updated chapters as well as new ones on neuroscience, exposure science, mental health, and forensic science. This edition of the manual has also gone through the thorough review process of the National Academy of Sciences.

As in previous editions, we continue to caution judges regarding the proper use of the reference guides. They are not intended to instruct judges concerning what evidence should be admissible or to establish minimum standards for acceptable scientific testimony. Rather, the guides can assist judges in identifying the issues most commonly in dispute in these selected areas and in reaching an informed and reasoned assessment concerning the basis of expert evidence. They are designed to facilitate the process of identifying and narrowing issues concerning scientific evidence by outlining for judges the pivotal issues in the areas of science that are often subject to dispute. Citations in the reference guides identify cases in which specific issues were raised; they are examples of other instances in which judges were faced with similar problems. By identifying scientific areas commonly in dispute, the guides should improve the quality of the dialogue between the judges and the parties concerning the basis of expert evidence.

In our committee discussions, we benefited from the judgment and wisdom of the many distinguished members of our committee, who gave time without compensation. They included Justice Ming Chin of the Supreme Court of California; Judge Pauline Newman of the U.S. Court of Appeals for the Federal Circuit in Washington, D.C.; Judge Kathleen MacDonald O'Malley of the U.S. Court of Appeals for the Federal Circuit; Judge Jed Rakoff of the U.S. District Court for the Southern District of New York; Channing Robertson, Ruth G. and William K. Bowes Professor, School of Engineering, and Professor, Department of Chemical Engineering, Stanford University; Joseph Rodricks,

*Reference Manual on Scientific Evidence*

Principal, Environ, Arlington, Virginia; Allen Wilcox, Senior Investigator, Institute of Environmental Health Sciences, Research Triangle Park, North Carolina; and Sandy Zabell, Professor of Statistics and Mathematics, Weinberg College of Arts and Sciences, Northwestern University.

Special commendation, however, goes to Anne-Marie Mazza, Director of the Committee on Science, Technology, and Law, and Joe Cecil of the Federal Judicial Center. These individuals not only shepherded each chapter and its revisions through the process, but provided critical advice on content and editing. They, not we, are the real editors.

Finally, we would like to express our gratitude for the superb assistance of Steven Kendall and for the diligent work of Guru Madhavan, Sara Maddox, Lillian Maloy, and Julie Phillips.

JEROME P. KASSIRER AND GLADYS KESSLER  
*Committee Co-Chairs*

# Summary Table of Contents

A detailed Table of Contents appears at the front of each chapter.

Introduction, 1	Stephen Breyer
The Admissibility of Expert Testimony, 11	Margaret A. Berger
How Science Works, 37	David Goodstein
Reference Guide on Forensic Identification Expertise, 55	Paul C. Giannelli, Edward J. Imwinkelried, & Joseph L. Peterson
Reference Guide on DNA Identification Evidence, 129	David H. Kaye & George Sensabaugh
Reference Guide on Statistics, 211	David H. Kaye & David A. Freedman
Reference Guide on Multiple Regression, 303	Daniel L. Rubinfeld
Reference Guide on Survey Research, 359	Shari Seidman Diamond
Reference Guide on Estimation of Economic Damages, 425	Mark A. Allen, Robert E. Hall, & Victoria A. Lazear
Reference Guide on Exposure Science, 503	Joseph V. Rodricks
Reference Guide on Epidemiology, 549	Michael D. Green, D. Michal Freedman, & Leon Gordis
Reference Guide on Toxicology, 633	Bernard D. Goldstein & Mary Sue Henifin
Reference Guide on Medical Testimony, 687	John B. Wong, Lawrence O. Gostin, & Oscar A. Cabrera
Reference Guide on Neuroscience, 747	Henry T. Greely & Anthony D. Wagner
Reference Guide on Mental Health Evidence, 813	Paul S. Appelbaum
Reference Guide on Engineering, 897	Channing R. Robertson, John E. Moalli, & David L. Black
Appendix A. Biographical Information of Committee and Staff, 961	



# Reference Guide on Statistics

DAVID H. KAYE AND DAVID A. FREEDMAN

*David H. Kaye, M.A., J.D., is Distinguished Professor of Law and Weiss Family Scholar, The Pennsylvania State University, University Park, and Regents' Professor Emeritus, Arizona State University Sandra Day O'Connor College of Law and School of Life Sciences, Tempe.*

*David A. Freedman, Ph.D., was Professor of Statistics, University of California, Berkeley.*

[Editor's Note: Sadly, Professor Freedman passed away during the production of this manual.]

## CONTENTS

- I. Introduction, 213
  - A. Admissibility and Weight of Statistical Studies, 214
  - B. Varieties and Limits of Statistical Expertise, 214
  - C. Procedures That Enhance Statistical Testimony, 215
    - 1. Maintaining professional autonomy, 215
    - 2. Disclosing other analyses, 216
    - 3. Disclosing data and analytical methods before trial, 216
- II. How Have the Data Been Collected? 216
  - A. Is the Study Designed to Investigate Causation? 217
    - 1. Types of studies, 217
    - 2. Randomized controlled experiments, 220
    - 3. Observational studies, 220
    - 4. Can the results be generalized? 222
  - B. Descriptive Surveys and Censuses, 223
    - 1. What method is used to select the units? 223
    - 2. Of the units selected, which are measured? 226
  - C. Individual Measurements, 227
    - 1. Is the measurement process reliable? 227
    - 2. Is the measurement process valid? 228
    - 3. Are the measurements recorded correctly? 229
  - D. What Is Random? 230
- III. How Have the Data Been Presented? 230
  - A. Are Rates or Percentages Properly Interpreted? 230
    - 1. Have appropriate benchmarks been provided? 230
    - 2. Have the data collection procedures changed? 231
    - 3. Are the categories appropriate? 231
    - 4. How big is the base of a percentage? 233
    - 5. What comparisons are made? 233
  - B. Is an Appropriate Measure of Association Used? 233

- C. Does a Graph Portray Data Fairly? 236
  - 1. How are trends displayed? 236
  - 2. How are distributions displayed? 236
- D. Is an Appropriate Measure Used for the Center of a Distribution? 238
- E. Is an Appropriate Measure of Variability Used? 239
- IV. What Inferences Can Be Drawn from the Data? 240
  - A. Estimation, 242
    - 1. What estimator should be used? 242
    - 2. What is the standard error? The confidence interval? 243
    - 3. How big should the sample be? 246
    - 4. What are the technical difficulties? 247
  - B. Significance Levels and Hypothesis Tests, 249
    - 1. What is the  $p$ -value? 249
    - 2. Is a difference statistically significant? 251
    - 3. Tests of interval estimates? 252
    - 4. Is the sample statistically significant? 253
  - C. Evaluating Hypothesis Tests, 253
    - 1. What is the power of the test? 253
    - 2. What about small samples? 254
    - 3. One tail or two? 255
    - 4. How many tests have been done? 256
    - 5. What are the rival hypotheses? 257
  - D. Posterior Probabilities, 258
- V. Correlation and Regression, 260
  - A. Scatter Diagrams, 260
  - B. Correlation Coefficients, 261
    - 1. Is the association linear? 262
    - 2. Do outliers influence the correlation coefficient? 262
    - 3. Does a confounding variable influence the coefficient? 262
  - C. Regression Lines, 264
    - 1. What are the slope and intercept? 265
    - 2. What is the unit of analysis? 266
  - D. Statistical Models, 268
- Appendix, 273
  - A. Frequentists and Bayesians, 273
  - B. The Spock Jury: Technical Details, 275
  - C. The Nixon Papers: Technical Details, 278
  - D. A Social Science Example of Regression: Gender Discrimination in Salaries, 279
    - 1. The regression model, 279
    - 2. Standard errors,  $t$ -statistics, and statistical significance, 281
- Glossary of Terms, 283
- References on Statistics, 302

# I. Introduction

Statistical assessments are prominent in many kinds of legal cases, including antitrust, employment discrimination, toxic torts, and voting rights cases.<sup>1</sup> This reference guide describes the elements of statistical reasoning. We hope the explanations will help judges and lawyers to understand statistical terminology, to see the strengths and weaknesses of statistical arguments, and to apply relevant legal doctrine. The guide is organized as follows:

- Section I provides an overview of the field, discusses the admissibility of statistical studies, and offers some suggestions about procedures that encourage the best use of statistical evidence.
- Section II addresses data collection and explains why the design of a study is the most important determinant of its quality. This section compares experiments with observational studies and surveys with censuses, indicating when the various kinds of study are likely to provide useful results.
- Section III discusses the art of summarizing data. This section considers the mean, median, and standard deviation. These are basic descriptive statistics, and most statistical analyses use them as building blocks. This section also discusses patterns in data that are brought out by graphs, percentages, and tables.
- Section IV describes the logic of statistical inference, emphasizing foundations and disclosing limitations. This section covers estimation, standard errors and confidence intervals, *p*-values, and hypothesis tests.
- Section V shows how associations can be described by scatter diagrams, correlation coefficients, and regression lines. Regression is often used to infer causation from association. This section explains the technique, indicating the circumstances under which it and other statistical models are likely to succeed—or fail.
- An appendix provides some technical details.
- The glossary defines statistical terms that may be encountered in litigation.

1. See generally *Statistical Science in the Courtroom* (Joseph L. Gastwirth ed., 2000); *Statistics and the Law* (Morris H. DeGroot et al. eds., 1986); National Research Council, *The Evolving Role of Statistical Assessments as Evidence in the Courts* (Stephen E. Fienberg ed., 1989) [hereinafter *The Evolving Role of Statistical Assessments as Evidence in the Courts*]; Michael O. Finkelstein & Bruce Levin, *Statistics for Lawyers* (2d ed. 2001); 1 & 2 Joseph L. Gastwirth, *Statistical Reasoning in Law and Public Policy* (1988); Hans Zeisel & David Kaye, *Prove It with Figures: Empirical Methods in Law and Litigation* (1997).



## A. Admissibility and Weight of Statistical Studies

Statistical studies suitably designed to address a material issue generally will be admissible under the Federal Rules of Evidence. The hearsay rule rarely is a serious barrier to the presentation of statistical studies, because such studies may be offered to explain the basis for an expert's opinion or may be admissible under the learned treatise exception to the hearsay rule.<sup>2</sup> Because most statistical methods relied on in court are described in textbooks or journal articles and are capable of producing useful results when properly applied, these methods generally satisfy important aspects of the "scientific knowledge" requirement in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*<sup>3</sup> Of course, a particular study may use a method that is entirely appropriate but that is so poorly executed that it should be inadmissible under Federal Rules of Evidence 403 and 702.<sup>4</sup> Or, the method may be inappropriate for the problem at hand and thus lack the "fit" spoken of in *Daubert*.<sup>5</sup> Or the study might rest on data of the type not reasonably relied on by statisticians or substantive experts and hence run afoul of Federal Rule of Evidence 703. Often, however, the battle over statistical evidence concerns weight or sufficiency rather than admissibility.

## B. Varieties and Limits of Statistical Expertise

For convenience, the field of statistics may be divided into three subfields: probability theory, theoretical statistics, and applied statistics. Probability theory is the mathematical study of outcomes that are governed, at least in part, by chance. Theoretical statistics is about the properties of statistical procedures, including error rates; probability theory plays a key role in this endeavor. Applied statistics draws on both of these fields to develop techniques for collecting or analyzing particular types of data.

2. See generally 2 McCormick on Evidence §§ 321, 324.3 (Kenneth S. Broun ed., 6th ed. 2006). Studies published by government agencies also may be admissible as public records. *Id.* § 296.

3. 509 U.S. 579, 589–90 (1993).

4. See *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 152 (1999) (suggesting that the trial court should "make certain that an expert, whether basing testimony upon professional studies or personal experience, employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field."); *Malletier v. Dooney & Bourke, Inc.*, 525 F. Supp. 2d 558, 562–63 (S.D.N.Y. 2007) ("While errors in a survey's methodology usually go to the weight accorded to the conclusions rather than its admissibility, . . . 'there will be occasions when the proffered survey is so flawed as to be completely unhelpful to the trier of fact.'") (quoting *AHP Subsidiary Holding Co. v. Stuart Hale Co.*, 1 F.3d 611, 618 (7th Cir.1993)).

5. *Daubert*, 509 U.S. at 591; *Anderson v. Westinghouse Savannah River Co.*, 406 F.3d 248 (4th Cir. 2005) (motion to exclude statistical analysis that compared black and white employees without adequately taking into account differences in their job titles or positions was properly granted under *Daubert*); *Malletier*, 525 F. Supp. 2d at 569 (excluding a consumer survey for "a lack of fit between the survey's questions and the law of dilution" and errors in the execution of the survey).

Statistical expertise is not confined to those with degrees in statistics. Because statistical reasoning underlies many kinds of empirical research, scholars in a variety of fields—including biology, economics, epidemiology, political science, and psychology—are exposed to statistical ideas, with an emphasis on the methods most important to the discipline.

Experts who specialize in using statistical methods, and whose professional careers demonstrate this orientation, are most likely to use appropriate procedures and correctly interpret the results. By contrast, forensic scientists often lack basic information about the studies underlying their testimony. *State v. Garrison*<sup>6</sup> illustrates the problem. In this murder prosecution involving bite mark evidence, a dentist was allowed to testify that “the probability factor of two sets of teeth being identical in a case similar to this is, approximately, eight in one million,” even though “he was unaware of the formula utilized to arrive at that figure other than that it was ‘computerized.’”<sup>7</sup>

At the same time, the choice of which data to examine, or how best to model a particular process, could require subject matter expertise that a statistician lacks. As a result, cases involving statistical evidence frequently are (or should be) “two expert” cases of interlocking testimony. A labor economist, for example, may supply a definition of the relevant labor market from which an employer draws its employees; the statistical expert may then compare the race of new hires to the racial composition of the labor market. Naturally, the value of the statistical analysis depends on the substantive knowledge that informs it.<sup>8</sup>

## C. Procedures That Enhance Statistical Testimony

### 1. Maintaining professional autonomy

Ideally, experts who conduct research in the context of litigation should proceed with the same objectivity that would be required in other contexts. Thus, experts who testify (or who supply results used in testimony) should conduct the analysis required to address in a professionally responsible fashion the issues posed by the litigation.<sup>9</sup> Questions about the freedom of inquiry accorded to testifying experts,

6. 585 P.2d 563 (Ariz. 1978).

7. *Id.* at 566, 568. For other examples, see David H. Kaye et al., *The New Wigmore: A Treatise on Evidence: Expert Evidence* § 12.2 (2d ed. 2011).

8. In *Vuyanich v. Republic National Bank*, 505 F. Supp. 224, 319 (N.D. Tex. 1980), *vacated*, 723 F.2d 1195 (5th Cir. 1984), defendant’s statistical expert criticized the plaintiffs’ statistical model for an implicit, but restrictive, assumption about male and female salaries. The district court trying the case accepted the model because the plaintiffs’ expert had a “very strong guess” about the assumption, and her expertise included labor economics as well as statistics. *Id.* It is doubtful, however, that economic knowledge sheds much light on the assumption, and it would have been simple to perform a less restrictive analysis.

9. See *The Evolving Role of Statistical Assessments as Evidence in the Courts*, *supra* note 1, at 164 (recommending that the expert be free to consult with colleagues who have not been retained

as well as the scope and depth of their investigations, may reveal some of the limitations to the testimony.

## 2. *Disclosing other analyses*

Statisticians analyze data using a variety of methods. There is much to be said for looking at the data in several ways. To permit a fair evaluation of the analysis that is eventually settled on, however, the testifying expert can be asked to explain how that approach was developed. According to some commentators, counsel who know of analyses that do not support the client's position should reveal them, rather than presenting only favorable results.<sup>10</sup>

## 3. *Disclosing data and analytical methods before trial*

The collection of data often is expensive and subject to errors and omissions. Moreover, careful exploration of the data can be time-consuming. To minimize debates at trial over the accuracy of data and the choice of analytical techniques, pretrial discovery procedures should be used, particularly with respect to the quality of the data and the method of analysis.<sup>11</sup>

# II. How Have the Data Been Collected?

The interpretation of data often depends on understanding “study design”—the plan for a statistical study and its implementation.<sup>12</sup> Different designs are suited to answering different questions. Also, flaws in the data can undermine any statistical analysis, and data quality is often determined by study design.

In many cases, statistical studies are used to show causation. Do food additives cause cancer? Does capital punishment deter crime? Would additional disclosures

by any party to the litigation and that the expert receive a letter of engagement providing for these and other safeguards).

10. *Id.* at 167; cf. William W. Schwarzer, *In Defense of “Automatic Disclosure in Discovery,”* 27 Ga. L. Rev. 655, 658–59 (1993) (“[T]he lawyer owes a duty to the court to make disclosure of core information.”). The National Research Council also recommends that “if a party gives statistical data to different experts for competing analyses, that fact be disclosed to the testifying expert, if any.” The Evolving Role of Statistical Assessments as Evidence in the Courts, *supra* note 1, at 167.

11. See The Special Comm. on Empirical Data in Legal Decision Making, Recommendations on Pretrial Proceedings in Cases with Voluminous Data, *reprinted in* The Evolving Role of Statistical Assessments as Evidence in the Courts, *supra* note 1, app. F; see also David H. Kaye, *Improving Legal Statistics*, 24 Law & Soc’y Rev. 1255 (1990).

12. For introductory treatments of data collection, see, for example, David Freedman et al., *Statistics* (4th ed. 2007); Darrell Huff, *How to Lie with Statistics* (1993); David S. Moore & William I. Notz, *Statistics: Concepts and Controversies* (6th ed. 2005); Hans Zeisel, *Say It with Figures* (6th ed. 1985); Zeisel & Kaye, *supra* note 1.

in a securities prospectus cause investors to behave differently? The design of studies to investigate causation is the first topic of this section.<sup>13</sup>

Sample data can be used to describe a population. The population is the whole class of units that are of interest; the sample is the set of units chosen for detailed study. Inferences from the part to the whole are justified when the sample is representative. Sampling is the second topic of this section.

Finally, the accuracy of the data will be considered. Because making and recording measurements is an error-prone activity, error rates should be assessed and the likely impact of errors considered. Data quality is the third topic of this section.

## A. Is the Study Designed to Investigate Causation?

### 1. Types of studies

When causation is the issue, anecdotal evidence can be brought to bear. So can observational studies or controlled experiments. Anecdotal reports may be of value, but they are ordinarily more helpful in generating lines of inquiry than in proving causation.<sup>14</sup> Observational studies can establish that one factor is associ-

13. See also Michael D. Green et al., Reference Guide on Epidemiology, Section V, in this manual; Joseph Rodricks, Reference Guide on Exposure Science, Section E, in this manual.

14. In medicine, evidence from clinical practice can be the starting point for discovery of cause-and-effect relationships. For examples, see David A. Freedman, *On Types of Scientific Enquiry*, in *The Oxford Handbook of Political Methodology* 300 (Janet M. Box-Steffensmeier et al. eds., 2008). Anecdotal evidence is rarely definitive, and some courts have suggested that attempts to infer causation from anecdotal reports are inadmissible as unsound methodology under *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993). See, e.g., *McClain v. Metabolife Int'l, Inc.*, 401 F.3d 1233, 1244 (11th Cir. 2005) (“simply because a person takes drugs and then suffers an injury does not show causation. Drawing such a conclusion from temporal relationships leads to the blunder of the *post hoc ergo propter hoc* fallacy.”); *In re Baycol Prods. Litig.*, 532 F. Supp. 2d 1029, 1039–40 (D. Minn. 2007) (excluding a meta-analysis based on reports to the Food and Drug Administration of adverse events); *Leblanc v. Chevron USA Inc.*, 513 F. Supp. 2d 641, 650 (E.D. La. 2007) (excluding plaintiffs’ experts’ opinions that benzene causes myelofibrosis because the causal hypothesis “that has been generated by case reports . . . has not been confirmed by the vast majority of epidemiologic studies of workers being exposed to benzene and more generally, petroleum products.”), *vacated*, 275 Fed. App’x. 319 (5th Cir. 2008) (remanding for consideration of newer government report on health effects of benzene); *cf. Matrixx Initiatives, Inc. v. Siracusano*, 131 S. Ct. 1309, 1321 (2011) (concluding that adverse event reports combined with other information could be of concern to a reasonable investor and therefore subject to a requirement of disclosure under SEC Rule 10b-5, but stating that “the mere existence of reports of adverse events . . . says nothing in and of itself about whether the drug is causing the adverse events”). Other courts are more open to “differential diagnoses” based primarily on timing. *E.g.*, *Best v. Lowe’s Home Ctrs., Inc.*, 563 F.3d 171 (6th Cir. 2009) (reversing the exclusion of a physician’s opinion that exposure to propenyl chloride caused a man to lose his sense of smell because of the timing in this one case and the physician’s inability to attribute the change to anything else); *Kaye et al.*, *supra* note 7, §§ 8.7.2 & 12.5.1. See also *Matrixx Initiatives*, *supra*, at 1322 (listing “a temporal relationship” in a single patient as one indication of “a reliable causal link”).

ated with another, but work is needed to bridge the gap between association and causation. Randomized controlled experiments are ideally suited for demonstrating causation.

Anecdotal evidence usually amounts to reports that events of one kind are followed by events of another kind. Typically, the reports are not even sufficient to show association, because there is no comparison group. For example, some children who live near power lines develop leukemia. Does exposure to electrical and magnetic fields cause this disease? The anecdotal evidence is not compelling because leukemia also occurs among children without exposure.<sup>15</sup> It is necessary to compare disease rates among those who are exposed and those who are not. If exposure causes the disease, the rate should be higher among the exposed and lower among the unexposed. That would be association.

The next issue is crucial: Exposed and unexposed people may differ in ways other than the exposure they have experienced. For example, children who live near power lines could come from poorer families and be more at risk from other environmental hazards. Such differences can create the appearance of a cause-and-effect relationship. Other differences can mask a real relationship. Cause-and-effect relationships often are quite subtle, and carefully designed studies are needed to draw valid conclusions.

An epidemiological classic makes the point. At one time, it was thought that lung cancer was caused by fumes from tarring the roads, because many lung cancer patients lived near roads that recently had been tarred. This is anecdotal evidence. But the argument is incomplete. For one thing, most people—whether exposed to asphalt fumes or unexposed—did not develop lung cancer. A comparison of rates was needed. The epidemiologists found that exposed persons and unexposed persons suffered from lung cancer at similar rates: Tar was probably not the causal agent. Exposure to cigarette smoke, however, turned out to be strongly associated with lung cancer. This study, in combination with later ones, made a compelling case that smoking cigarettes is the main cause of lung cancer.<sup>16</sup>

A good study design compares outcomes for subjects who are exposed to some factor (the treatment group) with outcomes for other subjects who are

15. See National Research Council, Committee on the Possible Effects of Electromagnetic Fields on Biologic Systems (1997); Zeisel & Kaye, *supra* note 1, at 66–67. There are problems in measuring exposure to electromagnetic fields, and results are inconsistent from one study to another. For such reasons, the epidemiological evidence for an effect on health is inconclusive. National Research Council, *supra*; Zeisel & Kaye, *supra*; Edward W. Campion, *Power Lines, Cancer, and Fear*, 337 New Eng. J. Med. 44 (1997) (editorial); Martha S. Linet et al., *Residential Exposure to Magnetic Fields and Acute Lymphoblastic Leukemia in Children*, 337 New Eng. J. Med. 1 (1997); Gary Taubes, *Magnetic Field-Cancer Link: Will It Rest in Peace?*, 277 Science 29 (1997) (quoting various epidemiologists).

16. Richard Doll & A. Bradford Hill, *A Study of the Aetiology of Carcinoma of the Lung*, 2 Brit. Med. J. 1271 (1952). This was a matched case-control study. Cohort studies soon followed. See Green et al., *supra* note 13. For a review of the evidence on causation, see 38 International Agency for Research on Cancer (IARC), World Health Org., IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans: Tobacco Smoking (1986).

not exposed (the control group). Now there is another important distinction to be made—that between controlled experiments and observational studies. In a controlled experiment, the investigators decide which subjects will be exposed and which subjects will go into the control group. In observational studies, by contrast, the subjects themselves choose their exposures. Because of self-selection, the treatment and control groups are likely to differ with respect to influential factors other than the one of primary interest. (These other factors are called lurking variables or confounding variables.)<sup>17</sup> With the health effects of power lines, family background is a possible confounder; so is exposure to other hazards. Many confounders have been proposed to explain the association between smoking and lung cancer, but careful epidemiological studies have ruled them out, one after the other.

Confounding remains a problem to reckon with, even for the best observational research. For example, women with herpes are more likely to develop cervical cancer than other women. Some investigators concluded that herpes caused cancer: In other words, they thought the association was causal. Later research showed that the primary cause of cervical cancer was human papilloma virus (HPV). Herpes was a marker of sexual activity. Women who had multiple sexual partners were more likely to be exposed not only to herpes but also to HPV. The association between herpes and cervical cancer was due to other variables.<sup>18</sup>

What are “variables?” In statistics, a variable is a characteristic of units in a study. With a study of people, the unit of analysis is the person. Typical variables include income (dollars per year) and educational level (years of schooling completed): These variables describe people. With a study of school districts, the unit of analysis is the district. Typical variables include average family income of district residents and average test scores of students in the district: These variables describe school districts.

When investigating a cause-and-effect relationship, the variable that represents the effect is called the dependent variable, because it depends on the causes. The variables that represent the causes are called independent variables. With a study of smoking and lung cancer, the independent variable would be smoking (e.g., number of cigarettes per day), and the dependent variable would mark the presence or absence of lung cancer. Dependent variables also are called outcome variables or response variables. Synonyms for independent variables are risk factors, predictors, and explanatory variables.

17. For example, a confounding variable may be correlated with the independent variable and act causally on the dependent variable. If the units being studied differ on the independent variable, they are also likely to differ on the confounder. The confounder—not the independent variable—could therefore be responsible for differences seen on the dependent variable.

18. For additional examples and further discussion, see Freedman et al., *supra* note 12, at 12–28, 150–52; David A. Freedman, *From Association to Causation: Some Remarks on the History of Statistics*, 14 Stat. Sci. 243 (1999). Some studies find that herpes is a “cofactor,” which increases risk among women who are also exposed to HPV. Only certain strains of HPV are carcinogenic.

## 2. Randomized controlled experiments

In randomized controlled experiments, investigators assign subjects to treatment or control groups at random. The groups are therefore likely to be comparable, except for the treatment. This minimizes the role of confounding. Minor imbalances will remain, due to the play of random chance; the likely effect on study results can be assessed by statistical techniques.<sup>19</sup> The bottom line is that causal inferences based on well-executed randomized experiments are generally more secure than inferences based on well-executed observational studies.

The following example should help bring the discussion together. Today, we know that taking aspirin helps prevent heart attacks. But initially, there was some controversy. People who take aspirin rarely have heart attacks. This is anecdotal evidence for a protective effect, but it proves almost nothing. After all, few people have frequent heart attacks, whether or not they take aspirin regularly. A good study compares heart attack rates for two groups: people who take aspirin (the treatment group) and people who do not (the controls). An observational study would be easy to do, but in such a study the aspirin-takers are likely to be different from the controls. Indeed, they are likely to be sicker—that is why they are taking aspirin. The study would be biased against finding a protective effect. Randomized experiments are harder to do, but they provide better evidence. It is the experiments that demonstrate a protective effect.<sup>20</sup>

In summary, data from a treatment group without a control group generally reveal very little and can be misleading. Comparisons are essential. If subjects are assigned to treatment and control groups at random, a difference in the outcomes between the two groups can usually be accepted, within the limits of statistical error (*infra* Section IV), as a good measure of the treatment effect. However, if the groups are created in any other way, differences that existed before treatment may contribute to differences in the outcomes or mask differences that otherwise would become manifest. Observational studies succeed to the extent that the treatment and control groups are comparable—apart from the treatment.

## 3. Observational studies

The bulk of the statistical studies seen in court are observational, not experimental. Take the question of whether capital punishment deters murder. To conduct a randomized controlled experiment, people would need to be assigned randomly to a treatment group or a control group. People in the treatment group would know they were subject to the death penalty for murder; the

19. Randomization of subjects to treatment or control groups puts statistical tests of significance on a secure footing. Freedman et al., *supra* note 12, at 503–22, 545–63; see *infra* Section IV.

20. In other instances, experiments have banished strongly held beliefs. *E.g.*, Scott M. Lippman et al., Effect of Selenium and Vitamin E on Risk of Prostate Cancer and Other Cancers: The Selenium and Vitamin E Cancer Prevention Trial (SELECT), 301 JAMA 39 (2009).

controls would know that they were exempt. Conducting such an experiment is not possible.

Many studies of the deterrent effect of the death penalty have been conducted, all observational, and some have attracted judicial attention. Researchers have catalogued differences in the incidence of murder in states with and without the death penalty and have analyzed changes in homicide rates and execution rates over the years. When reporting on such observational studies, investigators may speak of “control groups” (e.g., the states without capital punishment) or claim they are “controlling for” confounding variables by statistical methods.<sup>21</sup> However, association is not causation. The causal inferences that can be drawn from analysis of observational data—no matter how complex the statistical technique—usually rest on a foundation that is less secure than that provided by randomized controlled experiments.

That said, observational studies can be very useful. For example, there is strong observational evidence that smoking causes lung cancer (*supra* Section II.A.1). Generally, observational studies provide good evidence in the following circumstances:

- The association is seen in studies with different designs, on different kinds of subjects, and done by different research groups.<sup>22</sup> That reduces the chance that the association is due to a defect in one type of study, a peculiarity in one group of subjects, or the idiosyncrasies of one research group.
- The association holds when effects of confounding variables are taken into account by appropriate methods, for example, comparing smaller groups that are relatively homogeneous with respect to the confounders.<sup>23</sup>
- There is a plausible explanation for the effect of the independent variable; alternative explanations in terms of confounding should be less plausible than the proposed causal link.<sup>24</sup>

21. A procedure often used to control for confounding in observational studies is regression analysis. The underlying logic is described *infra* Section V.D and in Daniel L. Rubinfield, Reference Guide on Multiple Regression, Section II, in this manual. *But see* Richard A. Berk, Regression Analysis: A Constructive Critique (2004); Rethinking Social Inquiry: Diverse Tools, Shared Standards (Henry E. Brady & David Collier eds., 2004); David A. Freedman, Statistical Models: Theory and Practice (2005); David A. Freedman, *Oasis or Mirage*, Chance, Spring 2008, at 59.

22. For example, case-control studies are designed one way and cohort studies another, with many variations. *See, e.g.*, Leon Gordis, Epidemiology (4th ed. 2008); *supra* note 16.

23. The idea is to control for the influence of a confounder by stratification—making comparisons separately within groups for which the confounding variable is nearly constant and therefore has little influence over the variables of primary interest. For example, smokers are more likely to get lung cancer than nonsmokers. Age, gender, social class, and region of residence are all confounders, but controlling for such variables does not materially change the relationship between smoking and cancer rates. Furthermore, many different studies—of different types and on different populations—confirm the causal link. That is why most experts believe that smoking causes lung cancer and many other diseases. For a review of the literature, see International Agency for Research on Cancer, *supra* note 16.

24. A. Bradford Hill, *The Environment and Disease: Association or Causation?*, 58 Proc. Royal Soc’y Med. 295 (1965); Alfred S. Evans, Causation and Disease: A Chronological Journey 187 (1993). Plausibility, however, is a function of time and circumstances.



Thus, evidence for the causal link does not depend on observed associations alone.

Observational studies can produce legitimate disagreement among experts, and there is no mechanical procedure for resolving such differences of opinion. In the end, deciding whether associations are causal typically is not a matter of statistics alone, but also rests on scientific judgment. There are, however, some basic questions to ask when appraising causal inferences based on empirical studies:

- Was there a control group? Unless comparisons can be made, the study has little to say about causation.
- If there was a control group, how were subjects assigned to treatment or control: through a process under the control of the investigator (a controlled experiment) or through a process outside the control of the investigator (an observational study)?
- If the study was a controlled experiment, was the assignment made using a chance mechanism (randomization), or did it depend on the judgment of the investigator?

If the data came from an observational study or a nonrandomized controlled experiment,

- How did the subjects come to be in treatment or in control groups?
- Are the treatment and control groups comparable?
- If not, what adjustments were made to address confounding?
- Were the adjustments sensible and sufficient?<sup>25</sup>

#### 4. *Can the results be generalized?*

*Internal validity* is about the specifics of a particular study: Threats to internal validity include confounding and chance differences between treatment and control groups. *External validity* is about using a particular study or set of studies to reach more general conclusions. A careful randomized controlled experiment on a large but unrepresentative group of subjects will have high internal validity but low external validity.

Any study must be conducted on certain subjects, at certain times and places, and using certain treatments. To extrapolate from the conditions of a study to more general conditions raises questions of external validity. For example, studies suggest that definitions of insanity given to jurors influence decisions in cases of incest. Would the definitions have a similar effect in cases of murder? Other studies indicate that recidivism rates for ex-convicts are not affected by provid-

25. Many courts have noted the importance of confounding variables. *E.g.*, *People Who Care v. Rockford Bd. of Educ.*, 111 F.3d 528, 537–38 (7th Cir. 1997) (educational achievement); *Hollander v. Sandoz Pharms. Corp.*, 289 F.3d 1193, 1213 (10th Cir. 2002) (stroke); *In re Proportionality Review Project (II)*, 757 A.2d 168 (N.J. 2000) (capital sentences).

ing them with temporary financial support after release. Would similar results be obtained if conditions in the labor market were different?

Confidence in the appropriateness of an extrapolation cannot come from the experiment itself. It comes from knowledge about outside factors that would or would not affect the outcome.<sup>26</sup> Sometimes, several studies, each having different limitations, all point in the same direction. This is the case, for example, with studies indicating that jurors who approve of the death penalty are more likely to convict in a capital case.<sup>27</sup> Convergent results support the validity of generalizations.

## B. Descriptive Surveys and Censuses

We now turn to a second topic—choosing units for study. A census tries to measure some characteristic of every unit in a population. This is often impractical. Then investigators use sample surveys, which measure characteristics for only part of a population. The accuracy of the information collected in a census or survey depends on how the units are selected for study and how the measurements are made.<sup>28</sup>

### 1. What method is used to select the units?

By definition, a census seeks to measure some characteristic of every unit in a whole population. It may fall short of this goal, in which case one must ask

26. Such judgments are easiest in the physical and life sciences, but even here, there are problems. For example, it may be difficult to infer human responses to substances that affect animals. First, there are often inconsistencies across test species: A chemical may be carcinogenic in mice but not in rats. Extrapolation from rodents to humans is even more problematic. Second, to get measurable effects in animal experiments, chemicals are administered at very high doses. Results are extrapolated—using mathematical models—to the very low doses of concern in humans. However, there are many dose–response models to use and few grounds for choosing among them. Generally, different models produce radically different estimates of the “virtually safe dose” in humans. David A. Freedman & Hans Zeisel, *From Mouse to Man: The Quantitative Assessment of Cancer Risks*, 3 Stat. Sci. 3 (1988). For these reasons, many experts—and some courts in toxic tort cases—have concluded that evidence from animal experiments is generally insufficient by itself to establish causation. See, e.g., Bruce N. Ames et al., *The Causes and Prevention of Cancer*, 92 Proc. Nat’l Acad. Sci. USA 5258 (1995); National Research Council, *Science and Judgment in Risk Assessment* 59 (1994) (“There are reasons based on both biologic principles and empirical observations to support the hypothesis that many forms of biologic responses, including toxic responses, can be extrapolated across mammalian species, including *Homo sapiens*, but the scientific basis of such extrapolation is not established with sufficient rigor to allow broad and definitive generalizations to be made.”).

27. Phoebe C. Ellsworth, *Some Steps Between Attitudes and Verdicts*, in *Inside the Juror* 42, 46 (Reid Hastie ed., 1993). Nonetheless, in *Lockhart v. McCree*, 476 U.S. 162 (1986), the Supreme Court held that the exclusion of opponents of the death penalty in the guilt phase of a capital trial does not violate the constitutional requirement of an impartial jury.

28. See Shari Seidman Diamond, *Reference Guide on Survey Research*, Sections III, IV, in this manual.

whether the missing data are likely to differ in some systematic way from the data that are collected.<sup>29</sup> The methodological framework of a scientific survey is different. With probability methods, a sampling frame (i.e., an explicit list of units in the population) must be created. Individual units then are selected by an objective, well-defined chance procedure, and measurements are made on the sampled units.

To illustrate the idea of a sampling frame, suppose that a defendant in a criminal case seeks a change of venue: According to him, popular opinion is so adverse that it would be difficult to impanel an unbiased jury. To prove the state of popular opinion, the defendant commissions a survey. The relevant population consists of all persons in the jurisdiction who might be called for jury duty. The sampling frame is the list of all potential jurors, which is maintained by court officials and is made available to the defendant. In this hypothetical case, the fit between the sampling frame and the population would be excellent.

In other situations, the sampling frame is more problematic. In an obscenity case, for example, the defendant can offer a survey of community standards.<sup>30</sup> The population comprises all adults in the legally relevant district, but obtaining a full list of such people may not be possible. Suppose the survey is done by telephone, but cell phones are excluded from the sampling frame. (This is usual practice.) Suppose too that cell phone users, as a group, hold different opinions from landline users. In this second hypothetical, the poll is unlikely to reflect the opinions of the cell phone users, no matter how many individuals are sampled and no matter how carefully the interviewing is done.

Many surveys do not use probability methods. In commercial disputes involving trademarks or advertising, the population of all potential purchasers of a product is hard to identify. Pollsters may resort to an easily accessible subgroup of the population, for example, shoppers in a mall.<sup>31</sup> Such convenience samples may be biased by the interviewer's discretion in deciding whom to approach—a form of

29. The U.S. Decennial Census generally does not count everyone that it should, and it counts some people who should not be counted. There is evidence that net undercount is greater in some demographic groups than others. Supplemental studies may enable statisticians to adjust for errors and omissions, but the adjustments rest on uncertain assumptions. See Lawrence D. Brown et al., *Statistical Controversies in Census 2000*, 39 *Jurimetrics J.* 347 (2007); David A. Freedman & Kenneth W. Wachter, *Methods for Census 2000 and Statistical Adjustments*, in *Social Science Methodology* 232 (Steven Turner & William Outhwaite eds., 2007) (reviewing technical issues and litigation surrounding census adjustment in 1990 and 2000); 9 *Stat. Sci.* 458 (1994) (symposium presenting arguments for and against adjusting the 1990 census).

30. On the admissibility of such polls, see *State v. Midwest Pride IV, Inc.*, 721 N.E.2d 458 (Ohio Ct. App. 1998) (holding one such poll to have been properly excluded and collecting cases from other jurisdictions).

31. *E.g.*, *Smith v. Wal-Mart Stores, Inc.*, 537 F. Supp. 2d 1302, 1333 (N.D. Ga. 2008) (treating a small mall-intercept survey as entitled to much less weight than a survey based on a probability sample); *R.J. Reynolds Tobacco Co. v. Loew's Theatres, Inc.*, 511 F. Supp. 867, 876 (S.D.N.Y. 1980) (questioning the propriety of basing a “nationally projectable statistical percentage” on a suburban mall intercept study).

selection bias—and the refusal of some of those approached to participate—non-response bias (*infra* Section II.B.2). Selection bias is acute when constituents write their representatives, listeners call into radio talk shows, interest groups collect information from their members, or attorneys choose cases for trial.<sup>32</sup>

There are procedures that attempt to correct for selection bias. In quota sampling, for example, the interviewer is instructed to interview so many women, so many older people, so many ethnic minorities, and the like. But quotas still leave discretion to the interviewers in selecting members of each demographic group and therefore do not solve the problem of selection bias.<sup>33</sup>

Probability methods are designed to avoid selection bias. Once the population is reduced to a sampling frame, the units to be measured are selected by a lottery that gives each unit in the sampling frame a known, nonzero probability of being chosen. Random numbers leave no room for selection bias.<sup>34</sup> Such procedures are used to select individuals for jury duty. They also have been used to choose “bellwether” cases for representative trials to resolve issues in a large group of similar cases.<sup>35</sup>

32. *E.g.*, *Pittsburgh Press Club v. United States*, 579 F.2d 751, 759 (3d Cir. 1978) (tax-exempt club’s mail survey of its members to show little sponsorship of income-producing uses of facilities was held to be inadmissible hearsay because it “was neither objective, scientific, nor impartial”), *rev’d on other grounds*, 615 F.2d 600 (3d Cir. 1980). *Cf. In re Chevron U.S.A., Inc.*, 109 F.3d 1016 (5th Cir. 1997). In that case, the district court decided to try 30 cases to resolve common issues or to ascertain damages in 3000 claims arising from Chevron’s allegedly improper disposal of hazardous substances. The court asked the opposing parties to select 15 cases each. Selecting 30 extreme cases, however, is quite different from drawing a random sample of 30 cases. Thus, the court of appeals wrote that although random sampling would have been acceptable, the trial court could not use the results in the 30 extreme cases to resolve issues of fact or ascertain damages in the untried cases. *Id.* at 1020. Those cases, it warned, were “not cases calculated to represent the group of 3000 claimants.” *Id.* See *infra* note 35.

A well-known example of selection bias is the 1936 *Literary Digest* poll. After successfully predicting the winner of every U.S. presidential election since 1916, the *Digest* used the replies from 2.4 million respondents to predict that Alf Landon would win the popular vote, 57% to 43%. In fact, Franklin Roosevelt won by a landslide vote of 62% to 38%. See Freedman et al., *supra* note 12, at 334–35. The *Digest* was so far off, in part, because it chose names from telephone books, rosters of clubs and associations, city directories, lists of registered voters, and mail order listings. *Id.* at 335, A–20 n.6. In 1936, when only one household in four had a telephone, the people whose names appeared on such lists tended to be more affluent. Lists that overrepresented the affluent had worked well in earlier elections, when rich and poor voted along similar lines, but the bias in the sampling frame proved fatal when the Great Depression made economics a salient consideration for voters.

33. See Freedman et al., *supra* note 12, at 337–39.

34. In simple random sampling, units are drawn at random without replacement. In particular, each unit has the same probability of being chosen for the sample. *Id.* at 339–41. More complicated methods, such as stratified sampling and cluster sampling, have advantages in certain applications. In systematic sampling, every fifth, tenth, or hundredth (in mathematical jargon, every *n*th) unit in the sampling frame is selected. If the units are not in any special order, then systematic sampling is often comparable to simple random sampling.

35. *E.g.*, *In re Simon II Litig.*, 211 F.R.D. 86 (E.D.N.Y. 2002), *vacated*, 407 F.3d 125 (2d Cir. 2005), *dismissed*, 233 F.R.D. 123 (E.D.N.Y. 2006); *In re Estate of Marcus Human Rights Litig.*, 910

## 2. Of the units selected, which are measured?

Probability sampling ensures that within the limits of chance (*infra* Section IV), the sample will be representative of the sampling frame. The question remains regarding which units actually get measured. When documents are sampled for audit, all the selected ones can be examined, at least in principle. Human beings are less easily managed, and some will refuse to cooperate. Surveys should therefore report nonresponse rates. A large nonresponse rate warns of bias, although supplemental studies may establish that nonrespondents are similar to respondents with respect to characteristics of interest.<sup>36</sup>

In short, a good survey defines an appropriate population, uses a probability method for selecting the sample, has a high response rate, and gathers accurate information on the sample units. When these goals are met, the sample tends to be representative of the population. Data from the sample can be extrapolated

F. Supp. 1460 (D. Haw. 1995), *aff'd sub nom.* Hilao v. Estate of Marcos, 103 F.3d 767 (9th Cir. 1996); Cimino v. Raymark Indus., Inc., 751 F. Supp. 649 (E.D. Tex. 1990), *rev'd*, 151 F.3d 297 (5th Cir. 1998); *cf. In re Chevron U.S.A., Inc.*, 109 F.3d 1016 (5th Cir. 1997) (discussed *supra* note 32). Although trials in a suitable random sample of cases can produce reasonable estimates of average damages, the propriety of precluding individual trials raises questions of due process and the right to trial by jury. See Thomas E. Willging, Mass Torts Problems and Proposals: A Report to the Mass Torts Working Group (Fed. Judicial Ctr. 1999); *cf. Wal-Mart Stores, Inc. v. Dukes*, 131 S. Ct. 2541, 2560–61 (2011). The cases and the views of commentators are described more fully in David H. Kaye & David A. Freedman, *Statistical Proof*, in 1 *Modern Scientific Evidence: The Law and Science of Expert Testimony* § 6:16 (David L. Faigman et al. eds., 2009–2010).

36. For discussions of nonresponse rates and admissibility of surveys conducted for litigation, see *Johnson v. Big Lots Stores, Inc.*, 561 F. Supp. 2d 567 (E.D. La. 2008) (fair labor standards); *United States v. Dentsply Int'l, Inc.*, 277 F. Supp. 2d 387, 437 (D. Del. 2003), *rev'd on other grounds*, 399 F.3d 181 (3d Cir. 2005) (antitrust).

The 1936 *Literary Digest* election poll (*supra* note 32) illustrates the dangers in nonresponse. Only 24% of the 10 million people who received questionnaires returned them. Most of the respondents probably had strong views on the candidates and objected to President Roosevelt's economic program. This self-selection is likely to have biased the poll. Maurice C. Bryson, *The Literary Digest Poll: Making of a Statistical Myth*, 30 *Am. Statistician* 184 (1976); Freedman et al., *supra* note 12, at 335–36. Even when demographic characteristics of the sample match those of the population, caution is indicated. See David Streitfeld, *Shere Hite and the Trouble with Numbers*, 1 *Chance* 26 (1988); Chamont Wang, *Sense and Nonsense of Statistical Inference: Controversy, Misuse, and Subtlety* 174–76 (1993).

In *United States v. Gometz*, 730 F.2d 475, 478 (7th Cir. 1984) (en banc), the Seventh Circuit recognized that “a low rate of response to juror questionnaires could lead to the underrepresentation of a group that is entitled to be represented on the qualified jury wheel.” Nonetheless, the court held that under the Jury Selection and Service Act of 1968, 28 U.S.C. §§ 1861–1878 (1988), the clerk did not abuse his discretion by failing to take steps to increase a response rate of 30%. According to the court, “Congress wanted to make it possible for all qualified persons to serve on juries, which is different from forcing all qualified persons to be available for jury service.” *Gometz*, 730 F.2d at 480. Although it might “be a good thing to follow up on persons who do not respond to a jury questionnaire,” the court concluded that Congress “was not concerned with anything so esoteric as nonresponse bias.” *Id.* at 479, 482; *cf. In re United States*, 426 F.3d 1 (1st Cir. 2005) (reaching the same result with respect to underrepresentation of African Americans resulting in part from nonresponse bias).

to describe the characteristics of the population. Of course, surveys may be useful even if they fail to meet these criteria. But then, additional arguments are needed to justify the inferences.

### C. Individual Measurements

#### 1. Is the measurement process reliable?

Reliability and validity are two aspects of accuracy in measurement. In statistics, reliability refers to reproducibility of results.<sup>37</sup> A reliable measuring instrument returns consistent measurements. A scale, for example, is perfectly reliable if it reports the same weight for the same object time and again. It may not be accurate—it may always report a weight that is too high or one that is too low—but the perfectly reliable scale always reports the same weight for the same object. Its errors, if any, are systematic: They always point in the same direction.

Reliability can be ascertained by measuring the same quantity several times; the measurements must be made independently to avoid bias. Given independence, the correlation coefficient (*infra* Section V.B) between repeated measurements can be used as a measure of reliability. This is sometimes called a test-retest correlation or a reliability coefficient.

A courtroom example is DNA identification. An early method of identification required laboratories to determine the lengths of fragments of DNA. By making independent replicate measurements of the fragments, laboratories determined the likelihood that two measurements differed by specified amounts.<sup>38</sup> Such results were needed to decide whether a discrepancy between a crime sample and a suspect sample was sufficient to exclude the suspect.<sup>39</sup>

Coding provides another example. In many studies, descriptive information is obtained on the subjects. For statistical purposes, the information usually has to be reduced to numbers. The process of reducing information to numbers is called “coding,” and the reliability of the process should be evaluated. For example, in a study of death sentencing in Georgia, legally trained evaluators examined short summaries of cases and ranked them according to the defendant’s culpability.<sup>40</sup>

37. Courts often use “reliable” to mean “that which can be relied on” for some purpose, such as establishing probable cause or crediting a hearsay statement when the declarant is not produced for confrontation. *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579, 590 n.9 (1993), for example, distinguishes “evidentiary reliability” from reliability in the technical sense of giving consistent results. We use “reliability” to denote the latter.

38. See National Research Council, *The Evaluation of Forensic DNA Evidence* 139–41 (1996).

39. *Id.*; National Research Council, *DNA Technology in Forensic Science* 61–62 (1992). Current methods are discussed in David H. Kaye & George Sensabaugh, *Reference Guide on DNA Identification Evidence*, Section II, in this manual.

40. David C. Baldus et al., *Equal Justice and the Death Penalty: A Legal and Empirical Analysis* 49–50 (1990).

Two different aspects of reliability should be considered. First, the “within-observer variability” of judgments should be small—the same evaluator should rate essentially identical cases in similar ways. Second, the “between-observer variability” should be small—different evaluators should rate the same cases in essentially the same way.

## 2. *Is the measurement process valid?*

Reliability is necessary but not sufficient to ensure accuracy. In addition to reliability, validity is needed. A valid measuring instrument measures what it is supposed to. Thus, a polygraph measures certain physiological responses to stimuli, for example, in pulse rate or blood pressure. The measurements may be reliable. Nonetheless, the polygraph is not valid as a lie detector unless the measurements it makes are well correlated with lying.<sup>41</sup>

When there is an established way of measuring a variable, a new measurement process can be validated by comparison with the established one. Breathalyzer readings can be validated against alcohol levels found in blood samples. LSAT scores used for law school admissions can be validated against grades earned in law school. A common measure of validity is the correlation coefficient between the predictor and the criterion (e.g., test scores and later performance).<sup>42</sup>

Employment discrimination cases illustrate some of the difficulties. Thus, plaintiffs suing under Title VII of the Civil Rights Act may challenge an employment test that has a disparate impact on a protected group, and defendants may try to justify the use of a test as valid, reliable, and a business necessity.<sup>43</sup> For validation, the most appropriate criterion variable is clear enough: job performance. However, plaintiffs may then turn around and challenge the validity of performance ratings. For reliability, administering the test twice to the same group of people may be impractical. Even if repeated testing is practical, it may be statistically inadvisable, because subjects may learn something from the first round of testing that affects their scores on the second round. Such “practice effects” are likely to compromise the independence of the two measurements, and independence is needed to estimate reliability. Statisticians therefore use internal evidence

41. See *United States v. Henderson*, 409 F.3d 1293, 1303 (11th Cir. 2005) (“while the physical responses recorded by a polygraph machine may be tested, ‘there is no available data to prove that those specific responses are attributable to lying.’”); National Research Council, *The Polygraph and Lie Detection* (2003) (reviewing the scientific literature).

42. As the discussion of the correlation coefficient indicates, *infra* Section V.B, the closer the coefficient is to 1, the greater the validity. For a review of data on test reliability and validity, see Paul R. Sackett et al., *High-Stakes Testing in Higher Education and Employment: Appraising the Evidence for Validity and Fairness*, 63 *Am. Psychologist* 215 (2008).

43. See, e.g., *Washington v. Davis*, 426 U.S. 229, 252 (1976); *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 430–32 (1975); *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971); *Lanning v. S.E. Penn. Transp. Auth.*, 308 F.3d 286 (3d Cir. 2002).

from the test itself. For example, if scores on the first half of the test correlate well with scores from the second half, then that is evidence of reliability.

A further problem is that test-takers are likely to be a select group. The ones who get the jobs are even more highly selected. Generally, selection attenuates (weakens) the correlations. There are methods for using internal measures of reliability to estimate test-retest correlations; there are other methods that correct for attenuation. However, such methods depend on assumptions about the nature of the test and the procedures used to select the test-takers and are therefore open to challenge.<sup>44</sup>

### 3. *Are the measurements recorded correctly?*

Judging the adequacy of data collection involves an examination of the process by which measurements are taken. Are responses to interviews coded correctly? Do mistakes distort the results? How much data are missing? What was done to compensate for gaps in the data? These days, data are stored in computer files. Cross-checking the files against the original sources (e.g., paper records), at least on a sample basis, can be informative.

Data quality is a pervasive issue in litigation and in applied statistics more generally. A programmer moves a file from one computer to another, and half the data disappear. The definitions of crucial variables are lost in the sands of time. Values get corrupted: Social security numbers come to have eight digits instead of nine, and vehicle identification numbers fail the most elementary consistency checks. Everybody in the company, from the CEO to the rawest mailroom trainee, turns out to have been hired on the same day. Many of the residential customers have last names that indicate commercial activity (“Happy Valley Farriers”). These problems seem humdrum by comparison with those of reliability and validity, but—unless caught in time—they can be fatal to statistical arguments.<sup>45</sup>

44. See Thad Dunning & David A. Freedman, *Modeling Selection Effects*, in *Social Science Methodology* 225 (Steven Turner & William Outhwaite eds., 2007); Howard Wainer & David Thissen, *True Score Theory: The Traditional Method*, in *Test Scoring* 23 (David Thissen & Howard Wainer eds., 2001).

45. See, e.g., *Malletier v. Dooney & Bourke, Inc.*, 525 F. Supp. 2d 558, 630 (S.D.N.Y. 2007) (coding errors contributed “to the cumulative effect of the methodological errors” that warranted exclusion of a consumer confusion survey); *EEOC v. Sears, Roebuck & Co.*, 628 F. Supp. 1264, 1304, 1305 (N.D. Ill. 1986) (“[E]rrors in EEOC’s mechanical coding of information from applications in its hired and nonhired samples also make EEOC’s statistical analysis based on this data less reliable.” The EEOC “consistently coded prior experience in such a way that less experienced women are considered to have the same experience as more experienced men” and “has made so many general coding errors that its data base does not fairly reflect the characteristics of applicants for commission sales positions at Sears.”), *aff’d*, 839 F.2d 302 (7th Cir. 1988). But see *Dalley v. Mich. Blue Cross-Blue Shield, Inc.*, 612 F. Supp. 1444, 1456 (E.D. Mich. 1985) (“although plaintiffs show that there were some mistakes in coding, plaintiffs still fail to demonstrate that these errors were so generalized and so pervasive that the entire study is invalid.”).



### D. *What Is Random?*

In the law, a selection process sometimes is called “random,” provided that it does not exclude identifiable segments of the population. Statisticians use the term in a far more technical sense. For example, if we were to choose one person at random from a population, in the strict statistical sense, we would have to ensure that everybody in the population is chosen with exactly the same probability. With a randomized controlled experiment, subjects are assigned to treatment or control at random in the strict sense—by tossing coins, throwing dice, looking at tables of random numbers, or more commonly these days, by using a random number generator on a computer. The same rigorous definition applies to random sampling. It is randomness in the technical sense that provides assurance of unbiased estimates from a randomized controlled experiment or a probability sample. Randomness in the technical sense also justifies calculations of standard errors, confidence intervals, and *p*-values (*infra* Sections IV–V). Looser definitions of randomness are inadequate for statistical purposes.

## III. How Have the Data Been Presented?

After data have been collected, they should be presented in a way that makes them intelligible. Data can be summarized with a few numbers or with graphical displays. However, the wrong summary can mislead.<sup>46</sup> Section III.A discusses rates or percentages and provides some cautionary examples of misleading summaries, indicating the kinds of questions that might be considered when summaries are presented in court. Percentages are often used to demonstrate statistical association, which is the topic of Section III.B. Section III.C considers graphical summaries of data, while Sections III.D and III.E discuss some of the basic descriptive statistics that are likely to be encountered in litigation, including the mean, median, and standard deviation.

### A. *Are Rates or Percentages Properly Interpreted?*

#### 1. *Have appropriate benchmarks been provided?*

The selective presentation of numerical information is like quoting someone out of context. Is a fact that “over the past three years,” a particular index fund of large-cap stocks “gained a paltry 1.9% a year” indicative of poor management? Considering that “the average large-cap value fund has returned just 1.3% a year,”

46. See generally Freedman et al., *supra* note 12; Huff, *supra* note 12; Moore & Notz, *supra* note 12; Zeisel, *supra* note 12.

a growth rate of 1.9% is hardly an indictment.<sup>47</sup> In this example and many others, it is helpful to find a benchmark that puts the figures into perspective.

## 2. Have the data collection procedures changed?

Changes in the process of collecting data can create problems of interpretation. Statistics on crime provide many examples. The number of petty larcenies reported in Chicago more than doubled one year—not because of an abrupt crime wave, but because a new police commissioner introduced an improved reporting system.<sup>48</sup> For a time, police officials in Washington, D.C., “demonstrated” the success of a law-and-order campaign by valuing stolen goods at \$49, just below the \$50 threshold then used for inclusion in the Federal Bureau of Investigation’s Uniform Crime Reports.<sup>49</sup> Allegations of manipulation in the reporting of crime from one time period to another are legion.<sup>50</sup>

Changes in data collection procedures are by no means limited to crime statistics. Indeed, almost all series of numbers that cover many years are affected by changes in definitions and collection methods. When a study includes such time-series data, it is useful to inquire about changes and to look for any sudden jumps, which may signal such changes.

## 3. Are the categories appropriate?

Misleading summaries also can be produced by the choice of categories to be used for comparison. In *Philip Morris, Inc. v. Loew’s Theatres, Inc.*,<sup>51</sup> and *R.J. Reynolds Tobacco Co. v. Loew’s Theatres, Inc.*,<sup>52</sup> Philip Morris and R.J. Reynolds sought an injunction to stop the maker of Triumph low-tar cigarettes from running advertisements claiming that participants in a national taste test preferred Triumph to other brands. Plaintiffs alleged that claims that Triumph was a “national taste test winner” or Triumph “beats” other brands were false and misleading. An exhibit introduced by the defendant contained the data shown in Table 1.<sup>53</sup> Only  $14\% + 22\% = 36\%$  of the sample preferred Triumph to Merit, whereas

47. Paul J. Lim, *In a Downturn, Buy and Hold or Quit and Fold?*, N.Y. Times, July 27, 2008.

48. James P. Levine et al., *Criminal Justice in America: Law in Action* 99 (1986) (referring to a change from 1959 to 1960).

49. D. Seidman & M. Couzens, *Getting the Crime Rate Down: Political Pressure and Crime Reporting*, 8 Law & Soc’y Rev. 457 (1974).

50. Michael D. Maltz, *Missing UCR Data and Divergence of the NCVS and UCR Trends*, in *Understanding Crime Statistics: Revisiting the Divergence of the NCVS and UCR* 269, 280 (James P. Lynch & Lynn A. Addington eds., 2007) (citing newspaper reports in Boca Raton, Atlanta, New York, Philadelphia, Broward County (Florida), and Saint Louis); Michael Vasquez, *Miami Police: FBI: Crime Stats Accurate*, Miami Herald, May 1, 2008.

51. 511 F. Supp. 855 (S.D.N.Y. 1980).

52. 511 F. Supp. 867 (S.D.N.Y. 1980).

53. *Philip Morris*, 511 F. Supp. at 866.

29% + 11% = 40% preferred Merit to Triumph. By selectively combining categories, however, the defendant attempted to create a different impression. Because 24% found the brands to be about the same, and 36% preferred Triumph, the defendant claimed that a clear majority (36% + 24% = 60%) found Triumph “as good [as] or better than Merit.”<sup>54</sup> The court resisted this chicanery, finding that defendant’s test results did not support the advertising claims.<sup>55</sup>

Table 1. Data Used by a Defendant to Refute Plaintiffs’ False Advertising Claim

	Triumph Much Better Than Merit	Triumph Somewhat Better Than Merit	Triumph About the Same as Merit	Triumph Somewhat Worse Than Merit	Triumph Much Worse Than Merit
Number	45	73	77	93	36
Percentage	14	22	24	29	11

There was a similar distortion in claims for the accuracy of a home pregnancy test. The manufacturer advertised the test as 99.5% accurate under laboratory conditions. The data underlying this claim are summarized in Table 2.

Table 2. Home Pregnancy Test Results

	Actually Pregnant	Actually not Pregnant
Test says pregnant	197	0
Test says not pregnant	1	2
Total	198	2

Table 2 does indicate that only one error occurred in 200 assessments, or 99.5% overall accuracy, but the table also shows that the test can make two types of errors: It can tell a pregnant woman that she is not pregnant (a false negative), and it can tell a woman who is not pregnant that she is (a false positive). The reported 99.5% accuracy rate conceals a crucial fact—the company had virtually no data with which to measure the rate of false positives.<sup>56</sup>

54. *Id.*  
55. *Id.* at 856–57.  
56. Only two women in the sample were not pregnant; the test gave correct results for both of them. Although a false-positive rate of 0 is ideal, an estimate based on a sample of only two women is not. These data are reported in Arnold Barnett, *How Numbers Can Trick You*, Tech. Rev., Oct. 1994, at 38, 44–45.

#### 4. *How big is the base of a percentage?*

Rates and percentages often provide effective summaries of data, but these statistics can be misinterpreted. A percentage makes a comparison between two numbers: One number is the base, and the other number is compared to that base. Putting them on the same base (100) makes it easy to compare them.

When the base is small, however, a small change in absolute terms can generate a large percentage gain or loss. This could lead to newspaper headlines such as “Increase in Thefts Alarming,” even when the total number of thefts is small.<sup>57</sup> Conversely, a large base will make for small percentage increases. In these situations, actual numbers may be more revealing than percentages.

#### 5. *What comparisons are made?*

Finally, there is the issue of which numbers to compare. Researchers sometimes choose among alternative comparisons. It may be worthwhile to ask why they chose the one they did. Would another comparison give a different view? A government agency, for example, may want to compare the amount of service now being given with that of earlier years—but what earlier year should be the baseline? If the first year of operation is used, a large percentage increase should be expected because of startup problems. If last year is used as the base, was it also part of the trend, or was it an unusually poor year? If the base year is not representative of other years, the percentage may not portray the trend fairly. No single question can be formulated to detect such distortions, but it may help to ask for the numbers from which the percentages were obtained; asking about the base can also be helpful.<sup>58</sup>

### *B. Is an Appropriate Measure of Association Used?*

Many cases involve statistical association. Does a test for employee promotion have an exclusionary effect that depends on race or gender? Does the incidence of murder vary with the rate of executions for convicted murderers? Do consumer purchases of a product depend on the presence or absence of a product warning? This section discusses tables and percentage-based statistics that are frequently presented to answer such questions.<sup>59</sup>

Percentages often are used to describe the association between two variables. Suppose that a university alleged to discriminate against women in admitting

57. Lyda Longa, *Increase in Thefts Alarming*, Daytona News-J. June 8, 2008 (reporting a 35% increase in armed robberies in Daytona Beach, Florida, in a 5-month period, but not indicating whether the number had gone up by 6 (from 17 to 23), by 300 (from 850 to 1150), or by some other amount).

58. For assistance in coping with percentages, see Zeisel, *supra* note 12, at 1–24.

59. Correlation and regression are discussed *infra* Section V.

students consists of only two colleges—engineering and business. The university admits 350 out of 800 male applicants; by comparison, it admits only 200 out of 600 female applicants. Such data commonly are displayed as in Table 3.<sup>60</sup>

As Table 3 indicates,  $350/800 = 44\%$  of the males are admitted, compared with only  $200/600 = 33\%$  of the females. One way to express the disparity is to subtract the two percentages:  $44\% - 33\% = 11$  percentage points. Although such subtraction is commonly seen in jury discrimination cases,<sup>61</sup> the difference is inevitably small when the two percentages are both close to zero. If the selection rate for males is 5% and that for females is 1%, the difference is only 4 percentage points. Yet, females have only one-fifth the chance of males of being admitted, and that may be of real concern.

Table 3. Admissions by Gender

Decision	Male	Female	Total
Admit	350	200	550
Deny	450	400	850
Total	800	600	1400

For Table 3, the selection ratio (used by the Equal Employment Opportunity Commission in its “80% rule”) is  $33/44 = 75\%$ , meaning that, on average, women have 75% the chance of admission that men have.<sup>62</sup> However, the selection ratio has its own problems. In the last example, if the selection rates are 5% and 1%, then the exclusion rates are 95% and 99%. The ratio is  $99/95 = 104\%$ , meaning that females have, on average, 104% the risk of males of being rejected. The underlying facts are the same, of course, but this formulation sounds much less disturbing.

60. A table of this sort is called a “cross-tab” or a “contingency table.” Table 3 is “two-by-two” because it has two rows and two columns, not counting rows or columns containing totals.

61. See, e.g., *State v. Gibbs*, 758 A.2d 327, 337 (Conn. 2000); *Primeaux v. Dooley*, 747 N.W.2d 137, 141 (S.D. 2008); D.H. Kaye, *Statistical Evidence of Discrimination in Jury Selection*, in *Statistical Methods in Discrimination Litigation* 13 (David H. Kaye & Mikel Aickin eds., 1986).

62. A procedure that selects candidates from the least successful group at a rate less than 80% of the rate for the most successful group “will generally be regarded by the Federal enforcement agencies as evidence of adverse impact.” EEOC Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607.4(D) (2008). The rule is designed to help spot instances of substantially discriminatory practices, and the commission usually asks employers to justify any procedures that produce selection ratios of 80% or less.

The analogous statistic used in epidemiology is called the relative risk. See Green et al., *supra* note 13, Section III.A. Relative risks are usually quoted as decimals; for example, a selection ratio of 75% corresponds to a relative risk of 0.75.

The odds ratio is more symmetric. If 5% of male applicants are admitted, the odds on a man being admitted are  $5/95 = 1/19$ ; the odds on a woman being admitted are  $1/99$ . The odds ratio is  $(1/99)/(1/19) = 19/99$ . The odds ratio for rejection instead of acceptance is the same, except that the order is reversed.<sup>63</sup> Although the odds ratio has desirable mathematical properties, its meaning may be less clear than that of the selection ratio or the simple difference.

Data showing disparate impact are generally obtained by aggregating—putting together—statistics from a variety of sources. Unless the source material is fairly homogeneous, aggregation can distort patterns in the data. We illustrate the problem with the hypothetical admission data in Table 3. Applicants can be classified not only by gender and admission but also by the college to which they applied, as in Table 4.

Table 4. Admissions by Gender and College

Decision	Engineering		Business	
	Male	Female	Male	Female
Admit	300	100	50	100
Deny	300	100	150	300

The entries in Table 4 add up to the entries in Table 3. Expressed in a more technical manner, Table 3 is obtained by aggregating the data in Table 4. Yet there is no association between gender and admission in either college; men and women are admitted at identical rates. Combining two colleges with no association produces a university in which gender is associated strongly with admission. The explanation for this paradox is that the business college, to which most of the women applied, admits relatively few applicants. It is easier to be accepted at the engineering college, the college to which most of the men applied. This example illustrates a common issue: Association can result from combining heterogeneous statistical material.<sup>64</sup>

63. For women, the odds on rejection are 99 to 1; for men, 19 to 1. The ratio of these odds is 99/19. Likewise, the odds ratio for an admitted applicant being a man as opposed to a denied applicant being a man is also 99/19.

64. Tables 3 and 4 are hypothetical, but closely patterned on a real example. See P.J. Bickel et al., *Sex Bias in Graduate Admissions: Data from Berkeley*, 187 *Science* 398 (1975). The tables are an instance of Simpson's Paradox.

### C. Does a Graph Portray Data Fairly?

Graphs are useful for revealing key characteristics of a batch of numbers, trends over time, and the relationships among variables.

#### 1. How are trends displayed?

Graphs that plot values over time are useful for seeing trends. However, the scales on the axes matter. In Figure 1, the rate of all crimes of domestic violence in Florida (per 100,000 people) appears to decline rapidly over the 10 years from 1998 through 2007; in Figure 2, the same rate appears to drop slowly.<sup>65</sup> The moral is simple: Pay attention to the markings on the axes to determine whether the scale is appropriate.

Figure 1

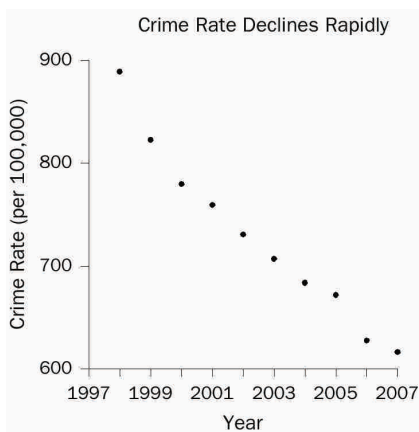
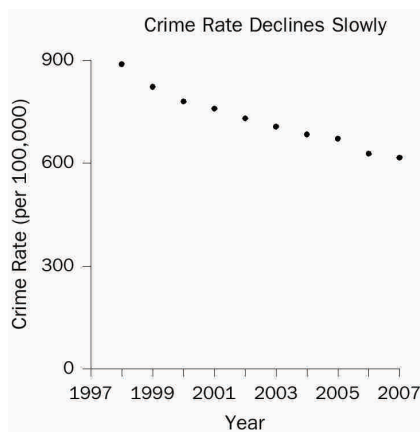


Figure 2



#### 2. How are distributions displayed?

A graph commonly used to display the distribution of data is the histogram. One axis denotes the numbers, and the other indicates how often those fall within

65. Florida Statistical Analysis Center, Florida Department of Law Enforcement, Florida's Crime Rate at a Glance, available at [http://www.fdle.state.fl.us/FSAC/Crime\\_Trends/domestic\\_violence/index.asp](http://www.fdle.state.fl.us/FSAC/Crime_Trends/domestic_violence/index.asp). The data are from the Florida Uniform Crime Report statistics on crimes ranging from simple stalking and forcible fondling to murder and arson. The Web page with the numbers graphed in Figures 1 and 2 is no longer posted, but similar data for all violent crime is available at [http://www.fdle.state.fl.us/FSAC/Crime\\_Trends/Violent-Crime.aspx](http://www.fdle.state.fl.us/FSAC/Crime_Trends/Violent-Crime.aspx).

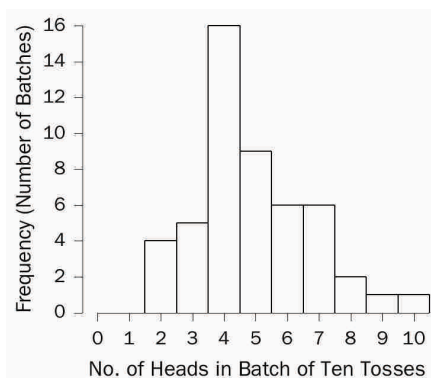
*Reference Guide on Statistics*

specified intervals (called “bins” or “class intervals”). For example, we flipped a quarter 10 times in a row and counted the number of heads in this “batch” of 10 tosses. With 50 batches, we obtained the following counts:<sup>66</sup>

7 7 5 6 8   4 2 3 6 5   4 3 4 7 4   6 8 4 7 4   7 4 5 4 3  
4 4 2 5 3   5 4 2 4 4   5 7 2 3 5   4 6 4 9 10   5 5 6 6 4

The histogram is shown in Figure 3.<sup>67</sup> A histogram shows how the data are distributed over the range of possible values. The spread can be made to appear larger or smaller, however, by changing the scale of the horizontal axis. Likewise, the shape can be altered somewhat by changing the size of the bins.<sup>68</sup> It may be worth inquiring how the analyst chose the bin widths.

Figure 3. Histogram showing how frequently various numbers of heads appeared in 50 batches of 10 tosses of a quarter.



66. The coin landed heads 7 times in the first 10 tosses; by coincidence, there were also 7 heads in the next 10 tosses; there were 5 heads in the third batch of 10 tosses; and so forth.

67. In Figure 3, the bin width is 1. There were no 0s or 1s in the data, so the bars over 0 and 1 disappear. There is a bin from 1.5 to 2.5; the four 2s in the data fall into this bin, so the bar over the interval from 1.5 to 2.5 has height 4. There is another bin from 2.5 to 3.5, which catches five 3s; the height of the corresponding bar is 5. And so forth.

All the bins in Figure 3 have the same width, so this histogram is just like a bar graph. However, data are often published in tables with unequal intervals. The resulting histograms will have unequal bin widths; bar heights should be calculated so that the areas (height  $\times$  width) are proportional to the frequencies. In general, a histogram differs from a bar graph in that it represents frequencies by area, not height. See Freedman et al., *supra* note 12, at 31–41.

68. As the width of the bins decreases, the graph becomes more detailed, but the appearance becomes more ragged until finally the graph is effectively a plot of each datum. The optimal bin width depends on the subject matter and the goal of the analysis.



### *D. Is an Appropriate Measure Used for the Center of a Distribution?*

Perhaps the most familiar descriptive statistic is the mean (or “arithmetic mean”). The mean can be found by adding all the numbers and dividing the total by how many numbers were added. By comparison, the median cuts the numbers into halves: half the numbers are larger than the median and half are smaller.<sup>69</sup> Yet a third statistic is the mode, which is the most common number in the dataset. These statistics are different, although they are not always clearly distinguished.<sup>70</sup> The mean takes account of all the data—it involves the total of all the numbers; however, particularly with small datasets, a few unusually large or small observations may have too much influence on the mean. The median is resistant to such outliers.

Thus, studies of damage awards in tort cases find that the mean is larger than the median.<sup>71</sup> This is because the mean takes into account (indeed, is heavily influenced by) the magnitudes of the relatively few very large awards, whereas the median merely counts their number. If one is seeking a single, representative number for the awards, the median may be more useful than the mean.<sup>72</sup> Still, if the issue is whether insurers were experiencing more costs from jury verdicts, the mean is the more appropriate statistic: The total of the awards is directly related to the mean, not to the median.<sup>73</sup>

69. Technically, at least half the numbers are at the median or larger; at least half are at the median or smaller. When the distribution is symmetric, the mean equals the median. The values diverge, however, when the distribution is asymmetric, or skewed.

70. In ordinary language, the arithmetic mean, the median, and the mode seem to be referred to interchangeably as “the average.” In statistical parlance, however, the average is the arithmetic mean. The mode is rarely used by statisticians, because it is unstable: Small changes to the data often result in large changes to the mode.

71. In a study using a probability sample of cases, the median compensatory award in wrongful death cases was \$961,000, whereas the mean award was around \$3.75 million for the 162 cases in which the plaintiff prevailed. Thomas H. Cohen & Steven K. Smith, U.S. Dep’t of Justice, Bureau of Justice Statistics Bulletin NCJ 202803, Civil Trial Cases and Verdicts in Large Counties 2001, 10 (2004). In *TXO Production Corp. v. Alliance Resources Corp.*, 509 U.S. 443 (1993), briefs portraying the punitive damage system as out of control pointed to mean punitive awards. These were some 10 times larger than the median awards described in briefs defending the system of punitive damages. Michael Rustad & Thomas Koenig, *The Supreme Court and Junk Social Science: Selective Distortion in Amicus Briefs*, 72 N.C. L. Rev. 91, 145–47 (1993).

72. In passing on proposed settlements in class-action lawsuits, courts have been advised to look to the magnitude of the settlements negotiated by the parties. But the mean settlement will be large if a higher number of meritorious, high-cost cases are resolved early in the life cycle of the litigation. This possibility led the court in *In re Educational Testing Service Praxis Principles of Learning and Teaching, Grades 7-12 Litig.*, 447 F. Supp. 2d 612, 625 (E.D. La. 2006), to regard the smaller median settlement as “more representative of the value of a typical claim than the mean value” and to use this median in extrapolating to the entire class of pending claims.

73. To get the total award, just multiply the mean by the number of awards; by contrast, the total cannot be computed from the median. (The more pertinent figure for the insurance industry is

Research also has shown that there is considerable stability in the ratio of punitive to compensatory damage awards, and the Supreme Court has placed great weight on this ratio in deciding whether punitive damages are excessive in a particular case. In *Exxon Shipping Co. v. Baker*,<sup>74</sup> Exxon contended that an award of \$2.5 billion in punitive damages for a catastrophic oil spill in Alaska was unreasonable under federal maritime law. The Court looked to a “comprehensive study of punitive damages awarded by juries in state civil trials [that] found a median ratio of punitive to compensatory awards of just 0.62:1, but a mean ratio of 2.90:1.”<sup>75</sup> The higher mean could reflect a relatively small but disturbing proportion of unjustifiably large punitive awards.<sup>76</sup> Looking to the median ratio as “the line near which cases like this one largely should be grouped,” the majority concluded that “a 1:1 ratio, which is above the median award, is a fair upper limit in such maritime cases [of reckless conduct].”<sup>77</sup>

### *E. Is an Appropriate Measure of Variability Used?*

The location of the center of a batch of numbers reveals nothing about the variations exhibited by these numbers.<sup>78</sup> Statistical measures of variability include the range, the interquartile range, and the standard deviation. The range is the difference between the largest number in the batch and the smallest. The range seems natural, and it indicates the maximum spread in the numbers, but the range is unstable because it depends entirely on the most extreme values.<sup>79</sup> The interquartile range is the difference between the 25th and 75th percentiles.<sup>80</sup> The interquartile range contains 50% of the numbers and is resistant to changes in extreme values. The standard deviation is a sort of mean deviation from the mean.<sup>81</sup>

not the total of jury awards, but actual claims experience including settlements; of course, even the risk of large punitive damage awards may have considerable impact.)

74. 128 S. Ct. 2605 (2008).

75. *Id.* at 2625.

76. According to the Court, “the outlier cases subject defendants to punitive damages that dwarf the corresponding compensatories,” and the “stark unpredictability” of these rare awards is the “real problem.” *Id.* This perceived unpredictability has been the subject of various statistical studies and much debate. See Anthony J. Sebok, *Punitive Damages: From Myth to Theory*, 92 Iowa L. Rev. 957 (2007).

77. 128 S. Ct. at 2633.

78. The numbers 1, 2, 5, 8, 9 have 5 as their mean and median. So do the numbers 5, 5, 5, 5, 5. In the first batch, the numbers vary considerably about their mean; in the second, the numbers do not vary at all.

79. Moreover, the range typically depends on the number of units in the sample.

80. By definition, 25% of the data fall below the 25th percentile, 90% fall below the 90th percentile, and so on. The median is the 50th percentile.

81. When the distribution follows the normal curve, about 68% of the data will be within 1 standard deviation of the mean, and about 95% will be within 2 standard deviations of the mean. For other distributions, the proportions will be different.

There are no hard and fast rules about which statistic is the best. In general, the bigger the measures of spread are, the more the numbers are dispersed.<sup>82</sup> Particularly in small datasets, the standard deviation can be influenced heavily by a few outlying values. To assess the extent of this influence, the mean and the standard deviation can be recomputed with the outliers discarded. Beyond this, any of the statistics can (and often should) be supplemented with a figure that displays much of the data.

## IV. What Inferences Can Be Drawn from the Data?

The inferences that may be drawn from a study depend on the design of the study and the quality of the data (*supra* Section II). The data might not address the issue of interest, might be systematically in error, or might be difficult to interpret because of confounding. Statisticians would group these concerns together under the rubric of “bias.” In this context, bias means systematic error, with no connotation of prejudice. We turn now to another concern, namely, the impact of random chance on study results (“random error”).<sup>83</sup>

If a pattern in the data is the result of chance, it is likely to wash out when more data are collected. By applying the laws of probability, a statistician can assess the likelihood that random error will create spurious patterns of certain kinds. Such assessments are often viewed as essential when making inferences from data.

Technically, the standard deviation is the square root of the variance; the variance is the mean square deviation from the mean. For example, if the mean is 100, then 120 deviates from the mean by 20, and the square of 20 is  $20^2 = 400$ . If the variance (i.e., the mean of the squared deviations) is 900, then the standard deviation is the square root of 900, that is,  $\sqrt{900} = 30$ . Taking the square root gets back to the original scale of the measurements. For example, if the measurements are of length in inches, the variance is in square inches; taking the square root changes back to inches.

82. In *Exxon Shipping Co. v. Baker*, 554 U.S. 471 (2008), along with the mean and median ratios of punitive to compensatory awards of 0.62 and 2.90, the Court referred to a standard deviation of 13.81. *Id.* at 498. These numbers led the Court to remark that “[e]ven to those of us unsophisticated in statistics, the thrust of these figures is clear: the spread is great, and the outlier cases subject defendants to punitive damages that dwarf the corresponding compensatories.” *Id.* at 499–500. The size of the standard deviation compared to the mean supports the observation that ratios in the cases of jury award studies are dispersed. A graph of each pair of punitive and compensatory damages offers more insight into how scattered these figures are. See Theodore Eisenberg et al., *The Predictability of Punitive Damages*, 26 J. Legal Stud. 623 (1997); *infra* Section V.A (explaining scatter diagrams).

83. Random error is also called sampling error, chance error, or statistical error. Econometricians use the parallel concept of random disturbance terms. See Rubinfeld, *supra* note 21. Randomness and cognate terms have precise technical meanings; it is randomness in the technical sense that justifies the probability calculations behind standard errors, confidence intervals, and *p*-values (*supra* Section II.D, *infra* Sections IV.A–B). For a discussion of samples and populations, see *supra* Section II.B.

Thus, statistical inference typically involves tasks such as the following, which will be discussed in the rest of this guide.

- *Estimation.* A statistician draws a sample from a population (*supra* Section II.B) and estimates a parameter—that is, a numerical characteristic of the population. (The average value of a large group of claims is a parameter of perennial interest.) Random error will throw the estimate off the mark. The question is, by how much? The precision of an estimate is usually reported in terms of the standard error and a confidence interval.
- *Significance testing.* A “null hypothesis” is formulated—for example, that a parameter takes a particular value. Because of random error, an estimated value for the parameter is likely to differ from the value specified by the null—even if the null is right. (“Null hypothesis” is often shortened to “null.”) How likely is it to get a difference as large as, or larger than, the one observed in the data? This chance is known as a *p*-value. Small *p*-values argue against the null hypothesis. Statistical significance is determined by reference to the *p*-value; significance testing (also called hypothesis testing) is the technique for computing *p*-values and determining statistical significance.
- *Developing a statistical model.* Statistical inferences often depend on the validity of statistical models for the data. If the data are collected on the basis of a probability sample or a randomized experiment, there will be statistical models that suit the occasion, and inferences based on these models will be secure. Otherwise, calculations are generally based on analogy: This group of people is like a random sample; that observational study is like a randomized experiment. The fit between the statistical model and the data collection process may then require examination—how good is the analogy? If the model breaks down, that will bias the analysis.
- *Computing posterior probabilities.* Given the sample data, what is the probability of the null hypothesis? The question might be of direct interest to the courts, especially when translated into English; for example, the null hypothesis might be the innocence of the defendant in a criminal case. Posterior probabilities can be computed using a formula called Bayes’ rule. However, the computation often depends on prior beliefs about the statistical model and its parameters; such prior beliefs almost necessarily require subjective judgment. According to the frequentist theory of statistics,<sup>84</sup>

84. The frequentist theory is also called objectivist, by contrast with the subjectivist version of Bayesian theory. In brief, frequentist methods treat probabilities as objective properties of the system being studied. Subjectivist Bayesians view probabilities as measuring subjective degrees of belief. See *infra* Section IV.D and Appendix, Section A, for discussion of the two positions. The Bayesian position is named after the Reverend Thomas Bayes (England, c. 1701–1761). His essay on the subject was published after his death: *An Essay Toward Solving a Problem in the Doctrine of Chances*, 53 Phil. Trans. Royal Soc’y London 370 (1763–1764). For discussion of the foundations and varieties of Bayesian and

prior probabilities rarely have meaning and neither do posterior probabilities.<sup>85</sup>

Key ideas of estimation and testing will be illustrated by courtroom examples, with some complications omitted for ease of presentation and some details postponed (*see infra* Section V.D on statistical models, and the Appendix on the calculations).

The first example, on estimation, concerns the Nixon papers. Under the Presidential Recordings and Materials Preservation Act of 1974, Congress impounded Nixon's presidential papers after he resigned. Nixon sued, seeking compensation on the theory that the materials belonged to him personally. Courts ruled in his favor: Nixon was entitled to the fair market value of the papers, with the amount to be proved at trial.<sup>86</sup>

The Nixon papers were stored in 20,000 boxes at the National Archives in Alexandria, Virginia. It was plainly impossible to value this entire population of material. Appraisers for the plaintiff therefore took a random sample of 500 boxes. (From this point on, details are simplified; thus, the example becomes somewhat hypothetical.) The appraisers determined the fair market value of each sample box. The average of the 500 sample values turned out to be \$2000. The standard deviation (*supra* Section III.E) of the 500 sample values was \$2200. Many boxes had low appraised values whereas some boxes were considered to be extremely valuable; this spread explains the large standard deviation.

## A. Estimation

### 1. What estimator should be used?

With the Nixon papers, it is natural to use the average value of the 500 sample boxes to estimate the average value of all 20,000 boxes comprising the population.

other forms of statistical inference, *see, e.g.*, Richard M. Royall, *Statistical Inference: A Likelihood Paradigm* (1997); James Berger, *The Case for Objective Bayesian Analysis*, 1 *Bayesian Analysis* 385 (2006), available at <http://ba.stat.cmu.edu/journal/2006/vol01/issue03/berger.pdf>; Stephen E. Fienberg, *Does It Make Sense to be an "Objective Bayesian"?* (Comment on Articles by Berger and by Goldstein), 1 *Bayesian Analysis* 429 (2006); David Freedman, *Some Issues in the Foundation of Statistics*, 1 *Found. Sci.* 19 (1995), reprinted in *Topics in the Foundation of Statistics* 19 (Bas C. van Fraassen ed., 1997); *see also* D.H. Kaye, *What Is Bayesianism?* in *Probability and Inference in the Law of Evidence: The Uses and Limits of Bayesianism* (Peter Tillers & Eric Green eds., 1988), reprinted in 28 *Jurimetrics J.* 161 (1988) (distinguishing between "Bayesian probability," "Bayesian statistical inference," "Bayesian inference writ large," and "Bayesian decision theory").

85. Prior probabilities of repeatable events (but not hypotheses) can be defined within the frequentist framework. *See infra* note 122. When this happens, prior and posterior probabilities for these events are meaningful according to both schools of thought.

86. *Nixon v. United States*, 978 F.2d 1269 (D.C. Cir. 1992); *Griffin v. United States*, 935 F. Supp. 1 (D.D.C. 1995).

With the average value for each box having been estimated as \$2000, the plaintiff demanded compensation in the amount of

$$20,000 \times \$2,000 = \$40,000,000.$$

In more complex problems, statisticians may have to choose among several estimators. Generally, estimators that tend to make smaller errors are preferred; however, “error” might be quantified in more than one way. Moreover, the advantage of one estimator over another may depend on features of the population that are largely unknown, at least before the data are collected and analyzed. For complicated problems, professional skill and judgment may therefore be required when choosing a sample design and an estimator. In such cases, the choices and the rationale for them should be documented.

## 2. What is the standard error? The confidence interval?

An estimate based on a sample is likely to be off the mark, at least by a small amount, because of random error. The standard error gives the likely magnitude of this random error, with smaller standard errors indicating better estimates.<sup>87</sup> In our example of the Nixon papers, the standard error for the sample average can be computed from (1) the size of the sample—500 boxes—and (2) the standard deviation of the sample values; *see infra* Appendix. Bigger samples give estimates that are more precise. Accordingly, the standard error should go down as the sample size grows, although the rate of improvement slows as the sample gets bigger. (“Sample size” and “the size of the sample” just mean the number of items in the sample; the “sample average” is the average value of the items in the sample.) The standard deviation of the sample comes into play by measuring heterogeneity. The less heterogeneity in the values, the smaller the standard error. For example, if all the values were about the same, a tiny sample would give an accurate estimate. Conversely, if the values are quite different from one another, a larger sample would be needed.

With a random sample of 500 boxes and a standard deviation of \$2200, the standard error for the sample average is about \$100. The plaintiff’s total demand was figured as the number of boxes (20,000) times the sample average (\$2000). Therefore, the standard error for the total demand can be computed as 20,000 times the standard error for the sample average<sup>88</sup>:

87. We distinguish between (1) the standard deviation of the sample, which measures the spread in the sample data and (2) the standard error of the sample average, which measures the likely size of the random error in the sample average. The standard error is often called the standard deviation, and courts generally use the latter term. *See, e.g., Castaneda v. Partida*, 430 U.S. 482 (1977).

88. We are assuming a simple random sample. Generally, the formula for the standard error must take into account the method used to draw the sample and the nature of the estimator. In fact, the Nixon appraisers used more elaborate statistical procedures. Moreover, they valued the material as of

$$20,000 \times \$100 = \$2,000,000.$$

How is the standard error to be interpreted? Just by the luck of the draw, a few too many high-value boxes may have come into the sample, in which case the estimate of \$40,000,000 is too high. Or, a few too many low-value boxes may have been drawn, in which case the estimate is too low. This is random error. The net effect of random error is unknown, because data are available only on the sample, not on the full population. However, the net effect is likely to be something close to the standard error of \$2,000,000. Random error throws the estimate off, one way or the other, by something close to the standard error. The role of the standard error is to gauge the likely size of the random error.

The plaintiff's argument may be open to a variety of objections, particularly regarding appraisal methods. However, the sampling plan is sound, as is the extrapolation from the sample to the population. And there is no need for a larger sample: The standard error is quite small relative to the total claim.

Random errors larger in magnitude than the standard error are commonplace. Random errors larger in magnitude than two or three times the standard error are unusual. Confidence intervals make these ideas more precise. Usually, a confidence interval for the population average is centered at the sample average; the desired confidence level is obtained by adding and subtracting a suitable multiple of the standard error. Statisticians who say that the population average falls within 1 standard error of the sample average will be correct about 68% of the time. Those who say "within 2 standard errors" will be correct about 95% of the time, and those who say "within 3 standard errors" will be correct about 99.7% of the time, and so forth. (We are assuming a large sample; the confidence levels correspond to areas under the normal curve and are approximations; the "population average" just means the average value of all the items in the population.<sup>89</sup>) In summary,

- To get a 68% confidence interval, start at the sample average, then add and subtract 1 standard error.
- To get a 95% confidence interval, start at the sample average, then add and subtract twice the standard error.

1995, extrapolated backward to the time of taking (1974), and then added interest. The text ignores these complications.

89. See *infra* Appendix. The area under the normal curve between  $-1$  and  $+1$  is close to 68.3%. Likewise, the area between  $-2$  and  $+2$  is close to 95.4%. Many academic statisticians would use  $\pm 1.96$  SE for a 95% confidence interval. However, the normal curve only gives an approximation to the relevant chances, and the error in that approximation will often be larger than a few tenths of a percent. For simplicity, we use  $\pm 1$  SE for the 68% confidence level, and  $\pm 2$  SE for 95% confidence. The normal curve gives good approximations when the sample size is reasonably large; for small samples, other techniques should be used. See *infra* notes 106–07.

*Reference Guide on Statistics*

- To get a 99.7% confidence interval, start at the sample average, then add and subtract three times the standard error.

With the Nixon papers, the 68% confidence interval for plaintiff's total demand runs

$$\begin{aligned} &\text{from } \$40,000,000 - \$2,000,000 = \$38,000,000 \\ &\text{to } \$40,000,000 + \$2,000,000 = \$42,000,000. \end{aligned}$$

The 95% confidence interval runs

$$\begin{aligned} &\text{from } \$40,000,000 - (2 \times \$2,000,000) = \$36,000,000 \\ &\text{to } \$40,000,000 + (2 \times \$2,000,000) = \$44,000,000. \end{aligned}$$

The 99.7% confidence interval runs

$$\begin{aligned} &\text{from } \$40,000,000 - (3 \times \$2,000,000) = \$34,000,000 \\ &\text{to } \$40,000,000 + (3 \times \$2,000,000) = \$46,000,000. \end{aligned}$$

To write this more compactly, we abbreviate standard error as SE. Thus, 1 SE is one standard error, 2 SE is twice the standard error, and so forth. With a large sample and an estimate like the sample average, a 68% confidence interval is the range

$$\text{estimate} - 1 \text{ SE to estimate} + 1 \text{ SE}.$$

A 95% confidence interval is the range

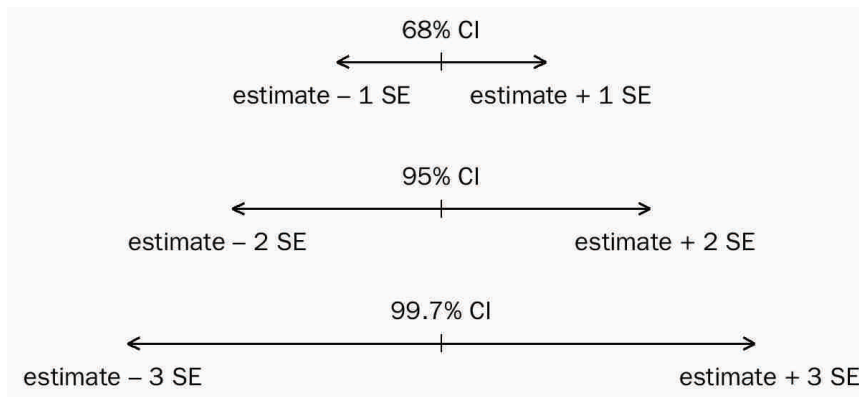
$$\text{estimate} - 2 \text{ SE to estimate} + 2 \text{ SE}.$$

The 99.7% confidence interval is the range

$$\text{estimate} - 3 \text{ SE to estimate} + 3 \text{ SE}.$$

For a given sample size, increased confidence can be attained only by widening the interval. The 95% confidence level is the most popular, but some authors use 99%, and 90% is seen on occasion. (The corresponding multipliers on the SE are about 2, 2.6, and 1.6, respectively; *see infra* Appendix.) The phrase “margin of error” generally means twice the standard error. In medical journals, “confidence interval” is often abbreviated as “CI.”





The main point is that an estimate based on a sample will differ from the exact population value, because of random error. The standard error gives the likely size of the random error. If the standard error is small, random error probably has little effect. If the standard error is large, the estimate may be seriously wrong. Confidence intervals are a technical refinement, and bias is a separate issue to consider (*infra* Section IV.A.4).

### 3. How big should the sample be?

There is no easy answer to this sensible question. Much depends on the level of error that is tolerable and the nature of the material being sampled. Generally, increasing the size of the sample will reduce the level of random error (“sampling error”). Bias (“nonsampling error”) cannot be reduced that way. Indeed, beyond some point, large samples are harder to manage and more vulnerable to non-sampling error. To reduce bias, the researcher must improve the design of the study or use a statistical model more tightly linked to the data collection process.

If the material being sampled is heterogeneous, random error will be large; a larger sample will be needed to offset the heterogeneity (*supra* Section IV.A.1). A pilot sample may be useful to estimate heterogeneity and determine the final sample size. Probability samples require some effort in the design phase, and it will rarely be sensible to draw a sample with fewer than, say, two or three dozen items. Moreover, with such small samples, methods based on the normal curve (*supra* Section IV.A.2) will not apply.

Population size (i.e., the number of items in the population) usually has little bearing on the precision of estimates for the population average. This is surprising. On the other hand, population size has a direct bearing on estimated totals. Both points are illustrated by the Nixon papers (*see supra* Section IV.A.2 and *infra* Appendix). To be sure, drawing a probability sample from a large population may

involve a lot of work. Samples presented in the courtroom have ranged from 5 (tiny) to 1.7 million (huge).<sup>90</sup>

#### 4. What are the technical difficulties?

To begin with, “confidence” is a term of art. The confidence level indicates the percentage of the time that intervals from repeated samples would cover the true value. The confidence level does not express the chance that repeated estimates would fall into the confidence interval.<sup>91</sup> With the Nixon papers, the 95% confidence interval should not be interpreted as saying that 95% of all random samples will produce estimates in the range from \$36 million to \$44 million. Moreover, the confidence level does not give the probability that the unknown parameter lies within the confidence interval.<sup>92</sup> For example, the 95% confidence level should not be translated to a 95% probability that the total value of the papers is in the range from \$36 million to \$44 million. According to the frequentist theory of statistics, probability statements cannot be made about population characteristics: Probability statements apply to the behavior of samples. That is why the different term “confidence” is used.

The next point to make is that for a given confidence level, a narrower interval indicates a more precise estimate, whereas a broader interval indicates less

90. See *Lebrilla v. Farmers Group, Inc.*, No. 00–CC–017185 (Cal. Super. Ct., Orange County, Dec. 5, 2006) (preliminary approval of settlement), a class action lawsuit on behalf of plaintiffs who were insured by Farmers and had automobile accidents. Plaintiffs alleged that replacement parts recommended by Farmers did not meet specifications: Small samples were used to evaluate these allegations. At the other extreme, it was proposed to adjust Census 2000 for undercount and overcount by reviewing a sample of 1.7 million persons. See Brown et al., *supra* note 29, at 353.

91. Opinions reflecting this misinterpretation include *In re Silicone Gel Breast Implants Prods. Liab. Litig.*, 318 F. Supp. 2d 879, 897 (C.D. Cal. 2004) (“a margin of error between 0.5 and 8.0 at the 95% confidence level . . . means that 95 times out of 100 a study of that type would yield a relative risk value somewhere between 0.5 and 8.0.”); *United States ex rel. Free v. Peters*, 806 F. Supp. 705, 713 n.6 (N.D. Ill. 1992) (“A 99% confidence interval, for instance, is an indication that if we repeated our measurement 100 times under identical conditions, 99 times out of 100 the point estimate derived from the repeated experimentation will fall within the initial interval estimate. . . .”), *rev’d in part*, 12 F.3d 700 (7th Cir. 1993). The more technically correct statement in the *Silicone Gel* case, for example, would be that “the confidence interval of 0.5 to 8.0 means that the relative risk in the population could fall within this wide range and that in roughly 95 times out of 100, random samples from the same population, the confidence intervals (however wide they might be) would include the population value (whatever it is).”

92. See, e.g., Freedman et al., *supra* note 12, at 383–86; *infra* Section IV.B.1. Consequently, it is misleading to suggest that “[a] 95% confidence interval means that there is a 95% probability that the ‘true’ relative risk falls within the interval” or that “the probability that the true value was . . . within two standard deviations of the mean . . . would be 95 percent.” *DeLuca v. Merrell Dow Pharms., Inc.*, 791 F. Supp. 1042, 1046 (D.N.J. 1992), *aff’d*, 6 F.3d 778 (3d Cir. 1993); *SmithKline Beecham Corp. v. Apotex Corp.*, 247 F. Supp. 2d 1011, 1037 (N.D. Ill. 2003), *aff’d on other grounds*, 403 F.3d 1331 (Fed. Cir. 2005).

precision.<sup>93</sup> A high confidence level with a broad interval means very little, but a high confidence level for a small interval is impressive, indicating that the random error in the sample estimate is low. For example, take a 95% confidence interval for a damage claim. An interval that runs from \$34 million to \$44 million is one thing, but –\$10 million to \$90 million is something else entirely. Statements about confidence without mention of an interval are practically meaningless.<sup>94</sup>

Standard errors and confidence intervals are often derived from statistical models for the process that generated the data. The model usually has parameters—numerical constants describing the population from which samples were drawn. When the values of the parameters are not known, the statistician must work backward, using the sample data to make estimates. That was the case here.<sup>95</sup> Generally, the chances needed for statistical inference are computed from a model and estimated parameter values.

If the data come from a probability sample or a randomized controlled experiment (*supra* Sections II.A–B), the statistical model may be connected tightly to the actual data collection process. In other situations, using the model may be tantamount to assuming that a sample of convenience is like a random sample, or that an observational study is like a randomized experiment. With the Nixon papers, the appraisers drew a random sample, and that justified the statistical

93. In *Cimino v. Raymark Industries, Inc.*, 751 F. Supp. 649 (E.D. Tex. 1990), *rev'd*, 151 F.3d 297 (5th Cir. 1998), the district court drew certain random samples from more than 6000 pending asbestos cases, tried these cases, and used the results to estimate the total award to be given to all plaintiffs in the pending cases. The court then held a hearing to determine whether the samples were large enough to provide accurate estimates. The court's expert, an educational psychologist, testified that the estimates were accurate because the samples matched the population on such characteristics as race and the percentage of plaintiffs still alive. *Id.* at 664. However, the matches occurred only in the sense that population characteristics fell within 99% confidence intervals computed from the samples. The court thought that matches within the 99% confidence intervals proved more than matches within 95% intervals. *Id.* This is backward. To be correct in a few instances with a 99% confidence interval is not very impressive—by definition, such intervals are broad enough to ensure coverage 99% of the time.

94. In *Hilao v. Estate of Marcos*, 103 F.3d 767 (9th Cir. 1996), for example, “an expert on statistics . . . testified that . . . a random sample of 137 claims would achieve ‘a 95% statistical probability that the same percentage determined to be valid among the examined claims would be applicable to the totality of [9541 facially valid] claims filed.’” *Id.* at 782. There is no 95% “statistical probability” that a percentage computed from a sample will be “applicable” to a population. One can compute a confidence interval from a random sample and be 95% confident that the interval covers some parameter. The computation can be done for a sample of virtually any size, with larger samples giving smaller intervals. What is missing from the opinion is a discussion of the widths of the relevant intervals. For the same reason, it is meaningless to testify, as an expert did in *Ayyad v. Sprint Spectrum, L.P.*, No. RG03-121510 (Cal. Super. Ct., Alameda County) (transcript, May 28, 2008, at 730), that a simple regression equation is trustworthy because the coefficient of the explanatory variable has “an extremely high indication of reliability to more than 99% confidence level.”

95. With the Nixon papers, one parameter is the average value of all 20,000 boxes, and another parameter is the standard deviation of the 20,000 values. These parameters can be used to approximate the distribution of the sample average. See *infra* Appendix. Regression models and their parameters are discussed *infra* Section V and in Rubinfeld, *supra* note 21.

calculations—if not the appraised values themselves. In many contexts, the choice of an appropriate statistical model is less than obvious. When a model does not fit the data collection process, estimates and standard errors will not be probative.

Standard errors and confidence intervals generally ignore systematic errors such as selection bias or nonresponse bias (*supra* Sections II.B.1–2). For example, after reviewing studies to see whether a particular drug caused birth defects, a court observed that mothers of children with birth defects may be more likely to remember taking a drug during pregnancy than mothers with normal children. This selective recall would bias comparisons between samples from the two groups of women. The standard error for the estimated difference in drug usage between the groups would ignore this bias, as would the confidence interval.<sup>96</sup>

## B. Significance Levels and Hypothesis Tests

### 1. What Is the *p*-value?

In 1969, Dr. Benjamin Spock came to trial in the U.S. District Court for Massachusetts. The charge was conspiracy to violate the Military Service Act. The jury was drawn from a panel of 350 persons selected by the clerk of the court. The panel included only 102 women—substantially less than 50%—although a majority of the eligible jurors in the community were female. The shortfall in women was especially poignant in this case: “Of all defendants, Dr. Spock, who had given wise and welcome advice on child-rearing to millions of mothers, would have liked women on his jury.”<sup>97</sup>

Can the shortfall in women be explained by the mere play of random chance? To approach the problem, a statistician would formulate and test a null hypothesis. Here, the null hypothesis says that the panel is like 350 persons drawn at random from a large population that is 50% female. The expected number of women drawn would then be 50% of 350, which is 175. The observed number of women is 102. The shortfall is  $175 - 102 = 73$ . How likely is it to find a disparity this large or larger, between observed and expected values? The probability is called *p*, or the *p*-value.

96. *Brock v. Merrell Dow Pharms., Inc.*, 874 F.2d 307, 311–12 (5th Cir.), *modified*, 884 F.2d 166 (5th Cir. 1989). In *Brock*, the court stated that the confidence interval took account of bias (in the form of selective recall) as well as random error. 874 F.2d at 311–12. This is wrong. Even if the sampling error were nonexistent—which would be the case if one could interview every woman who had a child during the period that the drug was available—selective recall would produce a difference in the percentages of reported drug exposure between mothers of children with birth defects and those with normal children. In this hypothetical situation, the standard error would vanish. Therefore, the standard error could disclose nothing about the impact of selective recall.

97. Hans Zeisel, *Dr. Spock and the Case of the Vanishing Women Jurors*, 37 U. Chi. L. Rev. 1 (1969). Zeisel’s reasoning was different from that presented in this text. The conviction was reversed on appeal without reaching the issue of jury selection. *United States v. Spock*, 416 F.2d 165 (1st Cir. 1965).

The  $p$ -value is the probability of getting data as extreme as, or more extreme than, the actual data—given that the null hypothesis is true. In the example,  $p$  turns out to be essentially zero. The discrepancy between the observed and the expected is far too large to explain by random chance. Indeed, even if the panel had included 155 women, the  $p$ -value would only be around 0.02, or 2%.<sup>98</sup> (If the population is more than 50% female,  $p$  will be even smaller.) In short, the jury panel was nothing like a random sample from the community.

Large  $p$ -values indicate that a disparity can easily be explained by the play of chance: The data fall within the range likely to be produced by chance variation. On the other hand, if  $p$  is very small, something other than chance must be involved: The data are far away from the values expected under the null hypothesis. Significance testing often seems to involve multiple negatives. This is because a statistical test is an argument by contradiction.

With the Dr. Spock example, the null hypothesis asserts that the jury panel is like a random sample from a population that is 50% female. The data contradict this null hypothesis because the disparity between what is observed and what is expected (according to the null) is too large to be explained as the product of random chance. In a typical jury discrimination case, small  $p$ -values help a defendant appealing a conviction by showing that the jury panel is not like a random sample from the relevant population; large  $p$ -values hurt. In the usual employment context, small  $p$ -values help plaintiffs who complain of discrimination—for example, by showing that a disparity in promotion rates is too large to be explained by chance; conversely, large  $p$ -values would be consistent with the defense argument that the disparity is just due to chance.

Because  $p$  is calculated by assuming that the null hypothesis is correct,  $p$  does not give the chance that the null is true. The  $p$ -value merely gives the chance of getting evidence against the null hypothesis as strong as or stronger than the evidence at hand. Chance affects the data, not the hypothesis. According to the frequency theory of statistics, there is no meaningful way to assign a numerical probability to the null hypothesis. The correct interpretation of the  $p$ -value can therefore be summarized in two lines:

$p$  is the probability of extreme data given the null hypothesis.  
 $p$  is not the probability of the null hypothesis given extreme data.<sup>99</sup>

98. With 102 women out of 350, the  $p$ -value is about  $2/10^{15}$ , where  $10^{15}$  is 1 followed by 15 zeros, that is, a quadrillion. See *infra* Appendix for the calculations.

99. Some opinions present a contrary view. *E.g.*, *Vasquez v. Hillery*, 474 U.S. 254, 259 n.3 (1986) (“the District Court . . . ultimately accepted . . . a probability of 2 in 1000 that the phenomenon was attributable to chance”); *Nat’l Abortion Fed. v. Ashcroft*, 330 F. Supp. 2d 436 (S.D.N.Y. 2004), *aff’d in part*, 437 F.3d 278 (2d Cir. 2006), *vacated*, 224 Fed. App’x. 88 (2d Cir. 2007) (“According to Dr. Howell, . . . a ‘P value’ of 0.30 . . . indicates that there is a thirty percent probability that the results of the . . . [s]tudy were merely due to chance alone.”). Such statements confuse the probability of the

To recapitulate the logic of significance testing: If  $p$  is small, the observed data are far from what is expected under the null hypothesis—too far to be readily explained by the operations of chance. That discredits the null hypothesis.

Computing  $p$ -values requires statistical expertise. Many methods are available, but only some will fit the occasion. Sometimes standard errors will be part of the analysis; other times they will not be. Sometimes a difference of two standard errors will imply a  $p$ -value of about 5%; other times it will not. In general, the  $p$ -value depends on the model, the size of the sample, and the sample statistics.

## 2. Is a difference statistically significant?

If an observed difference is in the middle of the distribution that would be expected under the null hypothesis, there is no surprise. The sample data are of the type that often would be seen when the null hypothesis is true. The difference is not significant, as statisticians say, and the null hypothesis cannot be rejected. On the other hand, if the sample difference is far from the expected value—according to the null hypothesis—then the sample is unusual. The difference is significant, and the null hypothesis is rejected. Statistical significance is determined by comparing  $p$  to a preset value, called the significance level.<sup>100</sup> The null hypothesis is rejected when  $p$  falls below this level.

In practice, statistical analysts typically use levels of 5% and 1%.<sup>101</sup> The 5% level is the most common in social science, and an analyst who speaks of significant results without specifying the threshold probably is using this figure. An unexplained reference to highly significant results probably means that  $p$  is less

kind of outcome observed, which is computed under some model of chance, with the probability that chance is the explanation for the outcome—the “transposition fallacy.”

Instances of the transposition fallacy in criminal cases are collected in David H. Kaye et al., *The New Wigmore: A Treatise on Evidence: Expert Evidence* §§ 12.8.2(b) & 14.1.2 (2d ed. 2011). In *McDaniel v. Brown*, 130 S. Ct. 665 (2010), for example, a DNA analyst suggested that a random match probability of 1/3,000,000 implied a .000033 probability that the DNA was not the source of the DNA found on the victim’s clothing. See David H. Kaye, “False But Highly Persuasive”: *How Wrong Were the Probability Estimates in McDaniel v. Brown?* 108 Mich. L. Rev. First Impressions 1 (2009).

100. Statisticians use the Greek letter alpha ( $\alpha$ ) to denote the significance level;  $\alpha$  gives the chance of getting a significant result, assuming that the null hypothesis is true. Thus,  $\alpha$  represents the chance of a false rejection of the null hypothesis (also called a false positive, a false alarm, or a Type I error). For example, suppose  $\alpha = 5\%$ . If investigators do many studies, and the null hypothesis happens to be true in each case, then about 5% of the time they would obtain significant results—and falsely reject the null hypothesis.

101. The Supreme Court implicitly referred to this practice in *Castaneda v. Partida*, 430 U.S. 482, 496 n.17 (1977), and *Hazelwood School District v. United States*, 433 U.S. 299, 311 n.17 (1977). In these footnotes, the Court described the null hypothesis as “suspect to a social scientist” when a statistic from “large samples” falls more than “two or three standard deviations” from its expected value under the null hypothesis. Although the Court did not say so, these differences produce  $p$ -values of about 5% and 0.3% when the statistic is normally distributed. The Court’s standard deviation is our standard error.

than 1%. These levels of 5% and 1% have become icons of science and the legal process. In truth, however, such levels are at best useful conventions.

Because the term “significant” is merely a label for a certain kind of  $p$ -value, significance is subject to the same limitations as the underlying  $p$ -value. Thus, significant differences may be evidence that something besides random error is at work. They are not evidence that this something is legally or practically important. Statisticians distinguish between statistical and practical significance to make the point. When practical significance is lacking—when the size of a disparity is negligible—there is no reason to worry about statistical significance.<sup>102</sup>

It is easy to mistake the  $p$ -value for the probability of the null hypothesis given the data (*supra* Section IV.B.1). Likewise, if results are significant at the 5% level, it is tempting to conclude that the null hypothesis has only a 5% chance of being correct.<sup>103</sup> This temptation should be resisted. From the frequentist perspective, statistical hypotheses are either true or false. Probabilities govern the samples, not the models and hypotheses. The significance level tells us what is likely to happen when the null hypothesis is correct; it does not tell us the probability that the hypothesis is true. Significance comes no closer to expressing the probability that the null hypothesis is true than does the underlying  $p$ -value.

### 3. Tests or interval estimates?

How can a highly significant difference be practically insignificant? The reason is simple:  $p$  depends not only on the magnitude of the effect, but also on the sample size (among other things). With a huge sample, even a tiny effect will be

102. *E.g.*, *Waisome v. Port Auth.*, 948 F.2d 1370, 1376 (2d Cir. 1991) (“though the disparity was found to be statistically significant, it was of limited magnitude.”); *United States v. Henderson*, 409 F.3d 1293, 1306 (11th Cir. 2005) (regardless of statistical significance, excluding law enforcement officers from jury service does not have a large enough impact on the composition of grand juries to violate the Jury Selection and Service Act); *cf. Thornburg v. Gingles*, 478 U.S. 30, 53–54 (1986) (repeating the district court’s explanation of why “the correlation between the race of the voter and the voter’s choice of certain candidates was [not only] statistically significant,” but also “so marked as to be substantively significant, in the sense that the results of the individual election would have been different depending upon whether it had been held among only the white voters or only the black voters.”).

103. *E.g.*, *Waisome*, 948 F.2d at 1376 (“Social scientists consider a finding of two standard deviations significant, meaning there is about one chance in 20 that the explanation for a deviation could be random . . . .”); *Adams v. Ameritech Serv., Inc.*, 231 F.3d 414, 424 (7th Cir. 2000) (“Two standard deviations is normally enough to show that it is extremely unlikely (. . . less than a 5% probability) that the disparity is due to chance”); *Magistrini v. One Hour Martinizing Dry Cleaning*, 180 F. Supp. 2d 584, 605 n.26 (D.N.J. 2002) (a “statistically significant . . . study shows that there is only 5% probability that an observed association is due to chance.”); *cf. Giles v. Wyeth, Inc.*, 500 F. Supp. 2d 1048, 1056 (S.D. Ill. 2007) (“While [plaintiff] admits that a  $p$ -value of .15 is three times higher than what scientists generally consider statistically significant—that is, a  $p$ -value of .05 or lower—she maintains that this “represents 85% certainty, which meets any conceivable concept of preponderance of the evidence.”).

highly significant.<sup>104</sup> For example, suppose that a company hires 52% of male job applicants and 49% of female applicants. With a large enough sample, a statistician could compute an impressively small  $p$ -value. This  $p$ -value would confirm that the difference does not result from chance, but it would not convert a trivial difference (52% versus 49%) into a substantial one.<sup>105</sup> In short, the  $p$ -value does not measure the strength or importance of an association.

A “significant” effect can be small. Conversely, an effect that is “not significant” can be large. By inquiring into the magnitude of an effect, courts can avoid being misled by  $p$ -values. To focus attention on more substantive concerns—the size of the effect and the precision of the statistical analysis—interval estimates (e.g., confidence intervals) may be more valuable than tests. Seeing a plausible range of values for the quantity of interest helps describe the statistical uncertainty in the estimate.

#### 4. Is the sample statistically significant?

Many a sample has been praised for its statistical significance or blamed for its lack thereof. Technically, this makes little sense. Statistical significance is about the difference between observations and expectations. Significance therefore applies to statistics computed from the sample, but not to the sample itself, and certainly not to the size of the sample. Findings can be statistically significant. Differences can be statistically significant (*supra* Section IV.B.2). Estimates can be statistically significant (*infra* Section V.D.2). By contrast, samples can be representative or unrepresentative. They can be chosen well or badly (*supra* Section II.B.1). They can be large enough to give reliable results or too small to bother with (*supra* Section IV.A.3). But samples cannot be “statistically significant,” if this technical phrase is to be used as statisticians use it.

### C. Evaluating Hypothesis Tests

#### 1. What is the power of the test?

When a  $p$ -value is high, findings are not significant, and the null hypothesis is not rejected. This could happen for at least two reasons:

104. See *supra* Section IV.B.2. Although some opinions seem to equate small  $p$ -values with “gross” or “substantial” disparities, most courts recognize the need to decide whether the underlying sample statistics reveal that a disparity is large. *E.g.*, *Washington v. People*, 186 P.3d 594 (Colo. 2008) (jury selection).

105. *Cf.* *Frazier v. Garrison Indep. Sch. Dist.*, 980 F.2d 1514, 1526 (5th Cir. 1993) (rejecting claims of intentional discrimination in the use of a teacher competency examination that resulted in retention rates exceeding 95% for all groups); *Washington*, 186 P.2d 594 (although a jury selection practice that reduced the representation of “African-Americans [from] 7.7 percent of the population [to] 7.4 percent of the county’s jury panels produced a highly statistically significant disparity, the small degree of exclusion was not constitutionally significant.”).



1. The null hypothesis is true.
2. The null is false—but, by chance, the data happened to be of the kind expected under the null.

If the power of a statistical study is low, the second explanation may be plausible. Power is the chance that a statistical test will declare an effect when there is an effect to be declared.<sup>106</sup> This chance depends on the size of the effect and the size of the sample. Discerning subtle differences requires large samples; small samples may fail to detect substantial differences.

When a study with low power fails to show a significant effect, the results may therefore be more fairly described as inconclusive than negative. The proof is weak because power is low. On the other hand, when studies have a good chance of detecting a meaningful association, failure to obtain significance can be persuasive evidence that there is nothing much to be found.<sup>107</sup>

## 2. What about small samples?

For simplicity, the examples of statistical inference discussed here (*supra* Sections IV.A–B) were based on large samples. Small samples also can provide useful

106. More precisely, power is the probability of rejecting the null hypothesis when the alternative hypothesis (*infra* Section IV.C.5) is right. Typically, this probability will depend on the values of unknown parameters, as well as the preset significance level  $\alpha$ . The power can be computed for any value of  $\alpha$  and any choice of parameters satisfying the alternative hypothesis. See *infra* Appendix for an example. Frequentist hypothesis testing keeps the risk of a false positive to a specified level (such as  $\alpha = 5\%$ ) and then tries to maximize power.

Statisticians usually denote power by the Greek letter beta ( $\beta$ ). However, some authors use  $\beta$  to denote the probability of *accepting* the null hypothesis when the alternative hypothesis is true; this usage is fairly standard in epidemiology. Accepting the null hypothesis when the alternative holds true is a false negative (also called a Type II error, a missed signal, or a false acceptance of the null hypothesis).

The chance of a false negative may be computed from the power. Some commentators have claimed that the cutoff for significance should be chosen to equalize the chance of a false positive and a false negative, on the ground that this criterion corresponds to the more-probable-than-not burden of proof. The argument is fallacious, because  $\alpha$  and  $\beta$  do not give the probabilities of the null and alternative hypotheses; see *supra* Sections IV.B.1–2; *supra* note 34. See also D.H. Kaye, *Hypothesis Testing in the Courtroom*, in *Contributions to the Theory and Application of Statistics: A Volume in Honor of Herbert Solomon* 331, 341–43 (Alan E. Gelfand ed., 1987).

107. Some formal procedures (meta-analysis) are available to aggregate results across studies. See, e.g., *In re Bextra and Celebrex Marketing Sales Practices and Prod. Liab. Litig.*, 524 F. Supp. 2d 1166, 1174, 1184 (N.D. Cal. 2007) (holding that “[a] meta-analysis of all available published and unpublished randomized clinical trials” of certain pain-relief medicine was admissible). In principle, the power of the collective results will be greater than the power of each study. However, these procedures have their own weakness. See, e.g., Richard A. Berk & David A. Freedman, *Statistical Assumptions as Empirical Commitments*, in *Punishment and Social Control: Essays in Honor of Sheldon Messinger* 235, 244–48 (T.G. Blomberg & S. Cohen eds., 2d ed. 2003); Michael Oakes, *Statistical Inference: A Commentary for the Social and Behavioral Sciences* (1986); Diana B. Petitti, *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis Methods for Quantitative Synthesis in Medicine* (2d ed. 2000).

information. Indeed, when confidence intervals and  $p$ -values can be computed, the interpretation is the same with small samples as with large ones.<sup>108</sup> The concern with small samples is not that they are beyond the ken of statistical theory, but that

1. The underlying assumptions are hard to validate.
2. Because approximations based on the normal curve generally cannot be used, confidence intervals may be difficult to compute for parameters of interest. Likewise,  $p$ -values may be difficult to compute for hypotheses of interest.<sup>109</sup>
3. Small samples may be unreliable, with large standard errors, broad confidence intervals, and tests having low power.

### 3. *One tail or two?*

In many cases, a statistical test can be done either one-tailed or two-tailed; the second method often produces a  $p$ -value twice as big as the first method. The methods are easily explained with a hypothetical example. Suppose we toss a coin 1000 times and get 532 heads. The null hypothesis to be tested asserts that the coin is fair. If the null is correct, the chance of getting 532 or more heads is 2.3%. That is a one-tailed test, whose  $p$ -value is 2.3%. To make a two-tailed test, the statistician computes the chance of getting 532 or more heads—or  $500 - 32 = 468$  heads or fewer. This is 4.6%. In other words, the two-tailed  $p$ -value is 4.6%. Because small  $p$ -values are evidence against the null hypothesis, the one-tailed test seems to produce stronger evidence than its two-tailed counterpart. However, the advantage is largely illusory, as the example suggests. (The two-tailed test may seem artificial, but it offers some protection against possible artifacts resulting from multiple testing—the topic of the next section.)

Some courts and commentators have argued for one or the other type of test, but a rigid rule is not required if significance levels are used as guidelines rather than as mechanical rules for statistical proof.<sup>110</sup> One-tailed tests often make it

108. Advocates sometimes contend that samples are “too small to allow for meaningful statistical analysis,” *United States v. New York City Bd. of Educ.*, 487 F. Supp. 2d 220, 229 (E.D.N.Y. 2007), and courts often look to the size of samples from earlier cases to determine whether the sample data before them are admissible or convincing. *Id.* at 230; *Timmerman v. U.S. Bank*, 483 F.3d 1106, 1116 n.4 (10th Cir. 2007). However, a meaningful statistical analysis yielding a significant result can be based on a small sample, and reliability does not depend on sample size alone (*see supra* Section IV.A.3, *infra* Section V.C.1). Well-known small-sample techniques include the sign test and Fisher’s exact test. *E.g.*, Michael O. Finkelstein & Bruce Levin, *Statistics for Lawyers* 154–56, 339–41 (2d ed. 2001); *see generally* E.L. Lehmann & H.J.M. d’Abrera, *Nonparametrics* (2d ed. 2006).

109. With large samples, approximate inferences (e.g., based on the central limit theorem, *see infra* Appendix) may be quite adequate. These approximations will not be satisfactory for small samples.

110. *See, e.g.*, *United States v. State of Delaware*, 93 Fair Empl. Prac. Cas. (BNA) 1248, 2004 WL 609331, \*10 n.4 (D. Del. 2004). According to formal statistical theory, the choice between one

easier to reach a threshold such as 5%, at least in terms of appearance. However, if we recognize that 5% is not a magic line, then the choice between one tail and two is less important—as long as the choice and its effect on the *p*-value are made explicit.

#### 4. How many tests have been done?

Repeated testing complicates the interpretation of significance levels. If enough comparisons are made, random error almost guarantees that some will yield “significant” findings, even when there is no real effect. To illustrate the point, consider the problem of deciding whether a coin is biased. The probability that a fair coin will produce 10 heads when tossed 10 times is  $(1/2)^{10} = 1/1024$ . Observing 10 heads in the first 10 tosses, therefore, would be strong evidence that the coin is biased. Nonetheless, if a fair coin is tossed a few thousand times, it is likely that at least one string of ten consecutive heads will appear. Ten heads in the first ten tosses means one thing; a run of ten heads somewhere along the way to a few thousand tosses of a coin means quite another. A test—looking for a run of ten heads—can be repeated too often.

Artifacts from multiple testing are commonplace. Because research that fails to uncover significance often is not published, reviews of the literature may produce an unduly large number of studies finding statistical significance.<sup>111</sup> Even a single researcher may examine so many different relationships that a few will achieve statistical significance by mere happenstance. Almost any large dataset—even pages from a table of random digits—will contain some unusual pattern that can be uncovered by diligent search. Having detected the pattern, the analyst can perform a statistical test for it, blandly ignoring the search effort. Statistical significance is bound to follow.

There are statistical methods for dealing with multiple looks at the data, which permit the calculation of meaningful *p*-values in certain cases.<sup>112</sup> However, no general solution is available, and the existing methods would be of little help in the typical case where analysts have tested and rejected a variety of models before arriving at the one considered the most satisfactory (see *infra* Section V on regression models). In these situations, courts should not be overly impressed with

tail or two can sometimes be made by considering the exact form of the alternative hypothesis (*infra* Section IV.C.5). But see Freedman et al., *supra* note 12, at 547–50. One-tailed tests at the 5% level are viewed as weak evidence—no weaker standard is commonly used in the technical literature. One-tailed tests are also called one-sided (with no pejorative intent); two-tailed tests are two-sided.

111. E.g., Philippa J. Easterbrook et al., *Publication Bias in Clinical Research*, 337 *Lancet* 867 (1991); John P.A. Ioannidis, *Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials*, 279 *JAMA* 281 (1998); Stuart J. Pocock et al., *Statistical Problems in the Reporting of Clinical Trials: A Survey of Three Medical Journals*, 317 *New Eng. J. Med.* 426 (1987).

112. See, e.g., Sandrine Dudoit & Mark J. van der Laan, *Multiple Testing Procedures with Applications to Genomics* (2008).

claims that estimates are significant. Instead, they should be asking how analysts developed their models.<sup>113</sup>

### 5. What are the rival hypotheses?

The  $p$ -value of a statistical test is computed on the basis of a model for the data: the null hypothesis. Usually, the test is made in order to argue for the alternative hypothesis: another model. However, on closer examination, both models may prove to be unreasonable. A small  $p$ -value means something is going on besides random error. The alternative hypothesis should be viewed as one possible explanation, out of many, for the data.

In *Mapes Casino, Inc. v. Maryland Casualty Co.*,<sup>114</sup> the court recognized the importance of explanations that the proponent of the statistical evidence had failed to consider. In this action to collect on an insurance policy, Mapes sought to quantify its loss from theft. It argued that employees were using an intermediary to cash in chips at other casinos. The casino established that over an 18-month period, the win percentage at its craps tables was 6%, compared to an expected value of 20%. The statistics proved that *something* was wrong at the craps tables—the discrepancy was too big to explain as the product of random chance. But the court was not convinced by plaintiff's alternative hypothesis. The court pointed to other possible explanations (Runyonesque activities such as skimming, scamming, and crossroading) that might have accounted for the discrepancy without implicating the suspect employees.<sup>115</sup> In short, rejection of the null hypothesis does not leave the proffered alternative hypothesis as the only viable explanation for the data.<sup>116</sup>

113. Intuition may suggest that the more variables included in the model, the better. However, this idea often turns out to be wrong. Complex models may reflect only accidental features of the data. Standard statistical tests offer little protection against this possibility when the analyst has tried a variety of models before settling on the final specification. See authorities cited, *supra* note 21.

114. 290 F. Supp. 186 (D. Nev. 1968).

115. *Id.* at 193. Skimming consists of “taking off the top before counting the drop,” scamming is “cheating by collusion between dealer and player,” and crossroading involves “professional cheaters among the players.” *Id.* In plainer language, the court seems to have ruled that the casino itself might be cheating, or there could have been cheaters other than the particular employees identified in the case. At the least, plaintiff's statistical evidence did not rule out such possibilities. Compare EEOC v. Sears, Roebuck & Co., 839 F.2d 302, 312 & n.9, 313 (7th Cir. 1988) (EEOC's regression studies showing significant differences did not establish liability because surveys and testimony supported the rival hypothesis that women generally had less interest in commission sales positions), with EEOC v. General Tel. Co., 885 F.2d 575 (9th Cir. 1989) (unsubstantiated rival hypothesis of “lack of interest” in “nontraditional” jobs insufficient to rebut prima facie case of gender discrimination); cf. *supra* Section II.A (problem of confounding).

116. *E.g.*, Coleman v. Quaker Oats Co., 232 F.3d 1271, 1283 (9th Cir. 2000) (a disparity with a  $p$ -value of “3 in 100 billion” did not demonstrate age discrimination because “Quaker never contends that the disparity occurred by chance, just that it did not occur for discriminatory reasons. When other pertinent variables were factored in, the statistical disparity diminished and finally disappeared.”).

### D. Posterior Probabilities

Standard errors,  $p$ -values, and significance tests are common techniques for assessing random error. These procedures rely on sample data and are justified in terms of the operating characteristics of statistical procedures.<sup>117</sup> However, frequentist statisticians generally will not compute the probability that a particular hypothesis is correct, given the data.<sup>118</sup> For example, a frequentist may postulate that a coin is fair: There is a 50-50 chance of landing heads, and successive tosses are independent. This is viewed as an empirical statement—potentially falsifiable—about the coin. It is easy to calculate the chance that a fair coin will turn up heads in the next 10 tosses: The answer (*see supra* Section IV.C.4) is  $1/1024$ . Therefore, observing 10 heads in a row brings into serious doubt the initial hypothesis of fairness.

But what of the converse probability: If the coin does land heads 10 times, what is the chance that it is fair?<sup>119</sup> To compute such converse probabilities, it is necessary to postulate initial probabilities that the coin is fair, as well as probabilities of unfairness to various degrees. In the frequentist theory of inference, such postulates are untenable: Probabilities are objective features of the situation that specify the chances of events or effects, not hypotheses or causes.

By contrast, in the Bayesian approach, probabilities represent subjective degrees of belief about hypotheses or causes rather than objective facts about observations. The observer must quantify beliefs about the chance that the coin is unfair to various degrees—in advance of seeing the data.<sup>120</sup> These subjective probabilities, like the probabilities governing the tosses of the coin, are set up to obey the axioms of probability theory. The probabilities for the various hypotheses about the coin, specified before data collection, are called prior probabilities.

117. Operating characteristics include the expected value and standard error of estimators, probabilities of error for statistical tests, and the like.

118. In speaking of “frequentist statisticians” or “Bayesian statisticians,” we do not mean to suggest that all statisticians fall on one side of the philosophical divide or the other. These are archetypes. Many practicing statisticians are pragmatists, using whatever procedure they think is appropriate for the occasion, and not concerning themselves greatly with what the numbers they obtain really mean.

119. We call this a converse probability because it is of the form  $P(H_0 | \text{data})$  rather than  $P(\text{data} | H_0)$ ; an equivalent phrase, “inverse probability,” also is used. Treating  $P(\text{data} | H_0)$  as if it were the converse probability  $P(H_0 | \text{data})$  is the transposition fallacy. For example, most U.S. senators are men, but few men are senators. Consequently, there is a high probability that an individual who is a senator is a man, but the probability that an individual who is a man is a senator is practically zero. For examples of the transposition fallacy in court opinions, see cases cited *supra* notes 98, 102. The frequentist  $p$ -value,  $P(\text{data} | H_0)$ , is generally not a good approximation to the Bayesian  $P(H_0 | \text{data})$ ; the latter includes considerations of power and base rates.

120. For example, let  $p$  be the unknown probability that the coin lands heads. What is the chance that  $p$  exceeds 0.1? 0.6? The Bayesian statistician must be prepared to answer such questions. Bayesian procedures are sometimes defended on the ground that the beliefs of any rational observer must conform to the Bayesian rules. However, the definition of “rational” is purely formal. *See* Peter C. Fishburn, *The Axioms of Subjective Probability*, 1 Stat. Sci. 335 (1986); Freedman, *supra* note 84; David Kaye, *The Laws of Probability and the Law of the Land*, 47 U. Chi. L. Rev. 34 (1979).

Prior probabilities can be updated, using Bayes' rule, given data on how the coin actually falls. (The Appendix explains the rule.) In short, a Bayesian statistician can compute posterior probabilities for various hypotheses about the coin, given the data. These posterior probabilities quantify the statistician's confidence in the hypothesis that a coin is fair.<sup>121</sup> Although such posterior probabilities relate directly to hypotheses of legal interest, they are necessarily subjective, for they reflect not just the data but also the subjective prior probabilities—that is, degrees of belief about hypotheses formulated prior to obtaining data.

Such analyses have rarely been used in court, and the question of their forensic value has been aired primarily in the academic literature. Some statisticians favor Bayesian methods, and some commentators have proposed using these methods in some kinds of cases.<sup>122</sup> The frequentist view of statistics is more conventional; subjective Bayesians are a well-established minority.<sup>123</sup>

121. Here, confidence has the meaning ordinarily ascribed to it, rather than the technical interpretation applicable to a frequentist confidence interval. Consequently, it can be related to the burden of persuasion. See D.H. Kaye, *Apples and Oranges: Confidence Coefficients and the Burden of Persuasion*, 73 Cornell L. Rev. 54 (1987).

122. See David H. Kaye et al., *The New Wigmore: A Treatise on Evidence: Expert Evidence* §§ 12.8.5, 14.3.2 (2d ed. 2010); David H. Kaye, *Rounding Up the Usual Suspects: A Legal and Logical Analysis of DNA Database Trawls*, 87 N.C. L. Rev. 425 (2009). In addition, as indicated in the Appendix, Bayes' rule is crucial in solving certain problems involving conditional probabilities of related events. For example, if the proportion of women with breast cancer in a region is known, along with the probability that a mammogram of an affected woman will be positive for cancer and that the mammogram of an unaffected woman will be negative, then one can compute the numbers of false-positive and false-negative mammography results that would be expected to arise in a population-wide screening program. Using Bayes' rule to diagnose a specific patient, however, is more problematic, because the prior probability that the patient has breast cancer may not equal the population proportion. Nevertheless, to overcome the tendency to focus on a test result without considering the "base rate" at which a condition occurs, a diagnostician can apply Bayes' rule to plausible base rates before making a diagnosis. Finally, Bayes' rule also is valuable as a device to explicate the meaning of concepts such as error rates, probative value, and transposition. See, e.g., David H. Kaye, *The Double Helix and the Law of Evidence* (2010); Wigmore, *supra*, § 7.3.2; David H. Kaye & Jonathan J. Koehler, *The Misquantification of Probative Value*, 27 Law & Hum. Behav. 645 (2003).

123. "Objective Bayesians" use Bayes' rule without eliciting prior probabilities from subjective beliefs. One strategy is to use preliminary data to estimate the prior probabilities and then apply Bayes' rule to that empirical distribution. This "empirical Bayes" procedure avoids the charge of subjectivism at the cost of departing from a fully Bayesian framework. With ample data, however, it can be effective and the estimates or inferences can be understood in frequentist terms. Another "objective" approach is to use "noninformative" priors that are supposed to be independent of all data and prior beliefs. However, the choice of such priors can be questioned, and the approach has been attacked by frequentists and subjective Bayesians. E.g., Joseph B. Kadane, *Is "Objective Bayesian Analysis" Objective, Bayesian, or Wise?*, 1 Bayesian Analysis 433 (2006), available at <http://ba.stat.cmu.edu/journal/2006/vol01/issue03/kadane.pdf>; Jon Williamson, *Philosophies of Probability*, in *Philosophy of Mathematics* 493 (Andrew Irvine ed., 2009) (discussing the challenges to objective Bayesianism).

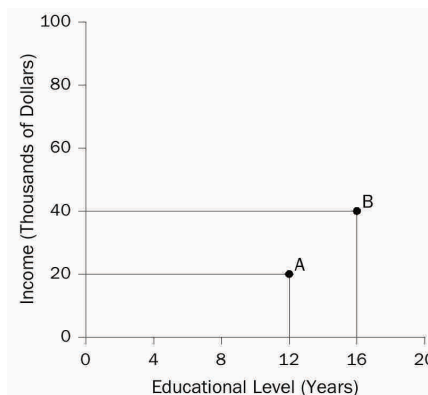
## V. Correlation and Regression

Regression models are used by many social scientists to infer causation from association. Such models have been offered in court to prove disparate impact in discrimination cases, to estimate damages in antitrust actions, and for many other purposes. Sections V.A, V.B, and V.C cover some preliminary material, showing how scatter diagrams, correlation coefficients, and regression lines can be used to summarize relationships between variables.<sup>124</sup> Section V.D explains the ideas and some of the pitfalls.

### A. Scatter Diagrams

The relationship between two variables can be graphed in a scatter diagram (also called a scatterplot or scattergram). We begin with data on income and education for a sample of 178 men, ages 25 to 34, residing in Kansas.<sup>125</sup> Each person in the sample corresponds to one dot in the diagram. As indicated in Figure 5, the horizontal axis shows education, and the vertical axis shows income. Person A completed 12 years of schooling (high school) and had an income of \$20,000. Person B completed 16 years of schooling (college) and had an income of \$40,000.

Figure 5. Plotting a scatter diagram. The horizontal axis shows educational level and the vertical axis shows income.

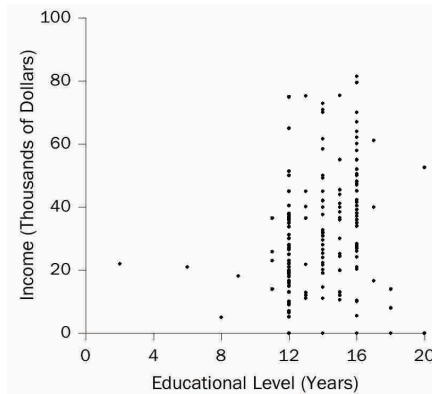


124. The focus is on simple linear regression. See also Rubinfeld, *supra* note 21, and the Appendix, *infra*, and Section II, *supra*, for further discussion of these ideas with an emphasis on econometrics.

125. These data are from a public-use CD, Bureau of the Census, U.S. Department of Commerce, for the March 2005 Current Population Survey. Income and education are self-reported. Income is censored at \$100,000. For additional details, see Freedman et al., *supra* note 12, at A-11. Both variables in a scatter diagram have to be quantitative (with numerical values) rather than qualitative (nonnumerical).

Figure 6 is the scatter diagram for the Kansas data. The diagram confirms an obvious point. There is a positive association between income and education. In general, persons with a higher educational level have higher incomes. However, there are many exceptions to this rule, and the association is not as strong as one might expect.

Figure 6. Scatter diagram for income and education: men ages 25 to 34 in Kansas.



## B. Correlation Coefficients

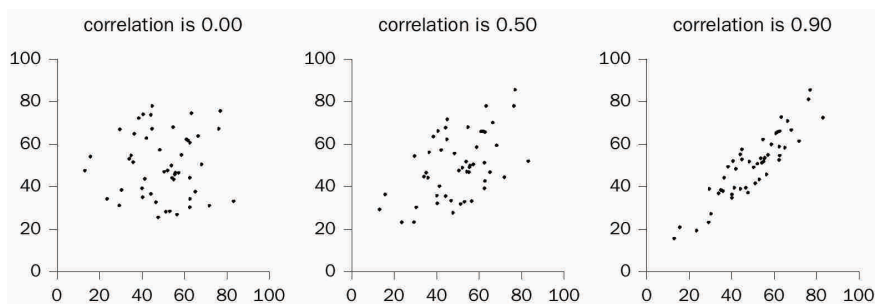
Two variables are positively correlated when their values tend to go up or down together, such as income and education in Figure 5. The correlation coefficient (usually denoted by the letter  $r$ ) is a single number that reflects the sign of an association and its strength. Figure 7 shows  $r$  for three scatter diagrams: In the first, there is no association; in the second, the association is positive and moderate; in the third, the association is positive and strong.

A correlation coefficient of 0 indicates no linear association between the variables. The maximum value for the coefficient is  $+1$ , indicating a perfect linear relationship: The dots in the scatter diagram fall on a straight line that slopes up. Sometimes, there is a negative association between two variables: Large values of one tend to go with small values of the other. The age of a car and its fuel economy in miles per gallon illustrate the idea. Negative association is indicated by negative values for  $r$ . The extreme case is an  $r$  of  $-1$ , indicating that all the points in the scatter diagram lie on a straight line that slopes down.

Weak associations are the rule in the social sciences. In Figure 5, the correlation between income and education is about 0.4. The correlation between college grades and first-year law school grades is under 0.3 at most law schools, while the



Figure 7. The correlation coefficient measures the sign of a linear association and its strength.



correlation between LSAT scores and first-year grades is generally about 0.4.<sup>126</sup> The correlation between heights of fraternal twins is about 0.5. By contrast, the correlation between heights of identical twins is about 0.95.

### 1. *Is the association linear?*

The correlation coefficient has a number of limitations, to be considered in turn. The correlation coefficient is designed to measure linear association. Figure 8 shows a strong nonlinear pattern with a correlation close to zero. The correlation coefficient is of limited use with nonlinear data.

### 2. *Do outliers influence the correlation coefficient?*

The correlation coefficient can be distorted by outliers—a few points that are far removed from the bulk of the data. The left-hand panel in Figure 9 shows that one outlier (lower right-hand corner) can reduce a perfect correlation to nearly nothing. Conversely, the right-hand panel shows that one outlier (upper right-hand corner) can raise a correlation of zero to nearly one. If there are extreme outliers in the data, the correlation coefficient is unlikely to be meaningful.

### 3. *Does a confounding variable influence the coefficient?*

The correlation coefficient measures the association between two variables. Researchers—and the courts—are usually more interested in causation. Causation is not the same as association. The association between two variables may be driven by a lurking variable that has been omitted from the analysis (*supra*

126. Lisa Anthony Stilwell et al., Predictive Validity of the LSAT: A National Summary of the 2001–2002 Correlation Studies 5, 8 (2003).

Reference Guide on Statistics

Figure 8. The scatter diagram shows a strong nonlinear association with a correlation coefficient close to zero. The correlation coefficient only measures the degree of linear association.

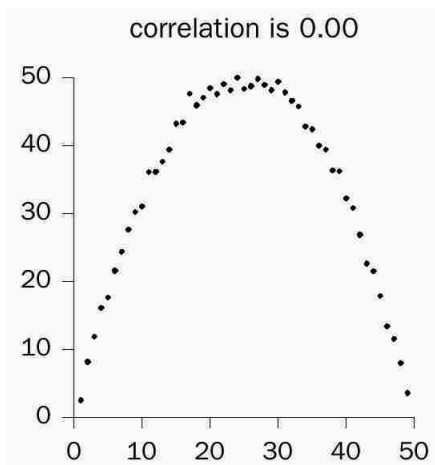
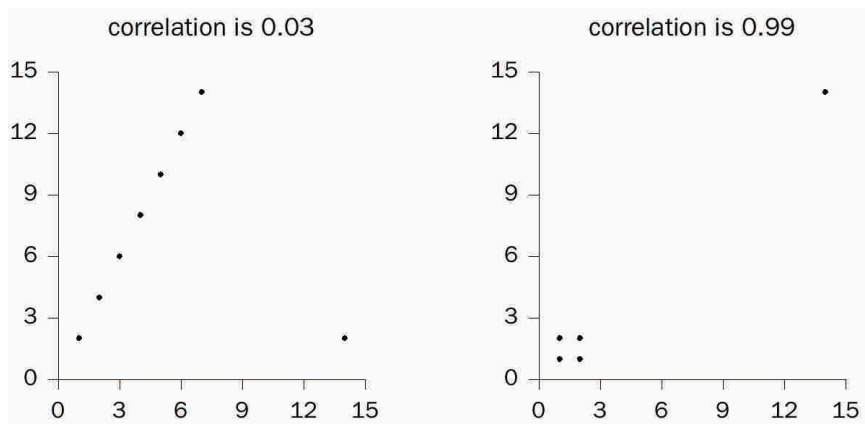


Figure 9. The correlation coefficient can be distorted by outliers.



Section II.A). For an easy example, there is an association between shoe size and vocabulary among schoolchildren. However, learning more words does not cause the feet to get bigger, and swollen feet do not make children more articulate. In this case, the lurking variable is easy to spot—age. In more realistic examples, the lurking variable is harder to identify.<sup>127</sup>

127. Green et al., *supra* note 13, Section IV.C, provides one such example.

In statistics, lurking variables are called confounders or confounding variables. Association often does reflect causation, but a large correlation coefficient is not enough to warrant causal inference. A large value of  $r$  only means that the dependent variable marches in step with the independent one: Possible reasons include causation, confounding, and coincidence. Multiple regression is one method that attempts to deal with confounders (*infra* Section V.D).<sup>128</sup>

### C. Regression Lines

The regression line can be used to describe a linear trend in the data. The regression line for income on education in the Kansas sample is shown in Figure 10. The height of the line estimates the average income for a given educational level. For example, the average income for people with 8 years of education is estimated at \$21,100, indicated by the height of the line at 8 years. The average income for people with 16 years of education is estimated at \$34,700.

Figure 10. The regression line for income on education and its estimates.

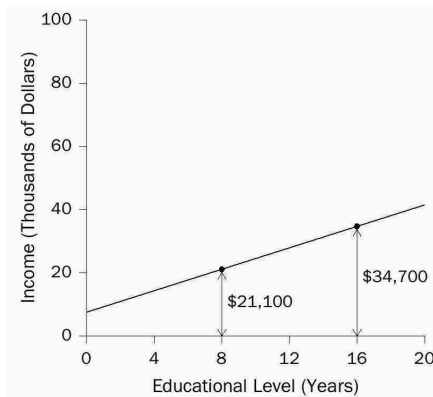
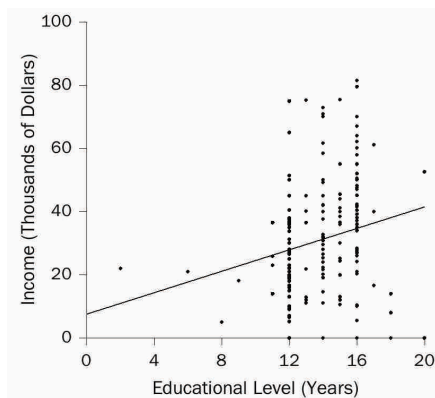


Figure 11 combines the data in Figures 5 and 10: it shows the scatter diagram for income and education, with the regression line superimposed. The line shows the average trend of income as education increases. Thus, the regression line indicates the extent to which a change in one variable (income) is associated with a change in another variable (education).

128. See also Rubinfeld, *supra* note 21. The difference between experiments and observational studies is discussed *supra* Section II.B.

Figure 11. Scatter diagram for income and education, with the regression line indicating the trend.



### 1. What are the slope and intercept?

The regression line can be described in terms of its intercept and slope. Often, the slope is the more interesting statistic. In Figure 11, the slope is \$1700 per year. On average, each additional year of education is associated with an additional \$1700 of income. Next, the intercept is \$7500. This is an estimate of the average income for (hypothetical) persons with zero years of education.<sup>129</sup> Figure 10 suggests this estimate may not be especially good. In general, estimates based on the regression line become less trustworthy as we move away from the bulk of the data.

The slope of the regression line has the same limitations as the correlation coefficient: (1) The slope may be misleading if the relationship is strongly non-linear and (2) the slope may be affected by confounders. With respect to (1), the slope of \$1700 per year in Figure 10 presents each additional year of education as having the same value, but some years of schooling surely are worth more and

129. The regression line, like any straight line, has an equation of the form  $y = a + bx$ . Here,  $a$  is the intercept (the value of  $y$  when  $x = 0$ ), and  $b$  is the slope (the change in  $y$  per unit change in  $x$ ). In Figure 9, the intercept of the regression line is \$7500 and the slope is \$1700 per year. The line estimates an average income of \$34,700 for people with 16 years of education. This may be computed from the intercept and slope as follows:

$$\$7500 + (\$1700 \text{ per year}) \times 16 \text{ years} = \$7500 + \$22,200 = \$34,700.$$

The slope  $b$  is the same anywhere along the line. Mathematically, that is what distinguishes straight lines from other curves. If the association is negative, the slope will be negative too. The slope is like the grade of a road, and it is negative if the road goes downhill. The intercept is like the starting elevation of a road, and it is computed from the data so that the line goes through the center of the scatter diagram, rather than being generally too high or too low.

others less. With respect to (2), the association between education and income is no doubt causal, but there are other factors to consider, including family background. Compared to individuals who did not graduate from high school, people with college degrees usually come from richer and better educated families. Thus, college graduates have advantages besides education. As statisticians might say, the effects of family background are confounded with the effects of education. Statisticians often use the guarded phrases “on average” and “associated with” when talking about the slope of the regression line. This is because the slope has limited utility when it comes to making causal inferences.

## 2. *What is the unit of analysis?*

If association between characteristics of individuals is of interest, these characteristics should be measured on individuals. Sometimes individual-level data are not to be had, but rates or averages for groups are available. “Ecological” correlations are computed from such rates or averages. These correlations generally overstate the strength of an association. For example, average income and average education can be determined for men living in each state and in Washington, D.C. The correlation coefficient for these 51 pairs of averages turns out to be 0.70. However, states do not go to school and do not earn incomes. People do. The correlation for income and education for men in the United States is only 0.42. The correlation for state averages overstates the correlation for individuals—a common tendency for ecological correlations.<sup>130</sup>

Ecological analysis is often seen in cases claiming dilution in voting strength of minorities. In this type of voting rights case, plaintiffs must prove three things: (1) the minority group constitutes a majority in at least one district of a proposed plan; (2) the minority group is politically cohesive, that is, votes fairly solidly for its preferred candidate; and (3) the majority group votes sufficiently as a bloc to defeat the minority-preferred candidate.<sup>131</sup> The first requirement is compactness; the second and third define polarized voting.

130. Correlations are computed from the March 2005 Current Population Survey for men ages 25–64. Freedman et al., *supra* note 12, at 149. The ecological correlation uses only the average figures, but within each state there is a lot of spread about the average. The ecological correlation smoothes away this individual variation. Cf. Green et al., *supra* note 13, Section II.B.4 (suggesting that ecological studies of exposure and disease are “far from conclusive” because of the lack of data on confounding variables (a much more general problem) as well as the possible aggregation bias described here); David A. Freedman, *Ecological Inference and the Ecological Fallacy*, in 6 Int’l Encyclopedia of the Social and Behavioral Sciences 4027 (Neil J. Smelser & Paul B. Baltes eds., 2001).

131. See *Thornburg v. Gingles*, 478 U.S. 30, 50–51 (1986) (“First, the minority group must be able to demonstrate that it is sufficiently large and geographically compact to constitute a majority in a single-member district. . . . Second, the minority group must be able to show that it is politically cohesive. . . . Third, the minority must be able to demonstrate that the white majority votes sufficiently as a bloc to enable it . . . usually to defeat the minority’s preferred candidate.”). In subsequent cases, the Court has emphasized that these factors are not sufficient to make out a violation of section 2 of

The secrecy of the ballot box means that polarized voting cannot be directly observed. Instead, plaintiffs in voting rights cases rely on ecological regression, with scatter diagrams, correlations, and regression lines to estimate voting behavior by groups and demonstrate polarization. The unit of analysis typically is the precinct. For each precinct, public records can be used to determine the percentage of registrants in each demographic group of interest, as well as the percentage of the total vote for each candidate—by voters from all demographic groups combined. Plaintiffs’ burden is to determine the vote by each demographic group separately.

Figure 12 shows how the argument unfolds. Each point in the scatter diagram represents data for one precinct in the 1982 Democratic primary election for auditor in Lee County, South Carolina. The horizontal axis shows the percentage of registrants who are white. The vertical axis shows the turnout rate for the white candidate. The regression line is plotted too. The slope would be interpreted as the difference between the white turnout rate and the black turnout rate for the white candidate. Furthermore, the intercept would be interpreted as the black turnout rate for the white candidate.<sup>132</sup> The validity of such estimates is contested in the statistical literature.<sup>133</sup>

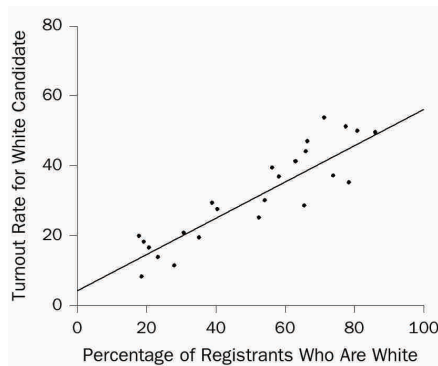
the Voting Rights Act. *E.g.*, *Johnson v. De Grandy*, 512 U.S. 997, 1011 (1994) (“*Gingles* . . . clearly declined to hold [these factors] sufficient in combination, either in the sense that a court’s examination of relevant circumstances was complete once the three factors were found to exist, or in the sense that the three in combination necessarily and in all circumstances demonstrated dilution.”).

132. By definition, the turnout rate equals the number of votes for the candidate, divided by the number of registrants; the rate is computed separately for each precinct. The intercept of the line in Figure 11 is 4%, and the slope is 0.52. Plaintiffs would conclude that only 4% of the black registrants voted for the white candidate, while  $4\% + 52\% = 56\%$  of the white registrants voted for the white candidate, which demonstrates polarization.

133. For further discussion of ecological regression in this context, see D. James Greiner, *Ecological Inference in Voting Rights Act Disputes: Where Are We Now, and Where Do We Want to Be?*, 47 *Jurimetrics J.* 115 (2007); Bernard Grofman & Chandler Davidson, *Controversies in Minority Voting: The Voting Rights Act in Perspective* (1992); Stephen P. Klein & David A. Freedman, *Ecological Regression in Voting Rights Cases*, 6 *Chance* 38 (Summer 1993). The use of ecological regression increased considerably after the Supreme Court noted in *Thornburg v. Gingles*, 478 U.S. 30, 53 n.20 (1986), that “[t]he District Court found both methods [extreme case analysis and bivariate ecological regression analysis] standard in the literature for the analysis of racially polarized voting.” See, e.g., *Cottier v. City of Martin*, 445 F.3d 1113, 1118 (8th Cir. 2006) (ecological regression is one of the “proven approaches to evaluating elections”); Bruce M. Clarke & Robert Timothy Reagan, *Fed. Judicial Ctr., Redistricting Litigation: An Overview of Legal, Statistical, and Case-Management Issues* (2002); Greiner, *supra*, at 117, 121. Nevertheless, courts have cautioned against “overreliance on bivariate ecological regression” in light of the inherent limitations of the technique. *Lewis v. Alamance County*, 99 F.3d 600, 604 n.3 (4th Cir. 1996); *Johnson v. Hamrick*, 296 F.3d 1065, 1080 n.4 (11th Cir. 2002) (“as a general rule, homogenous precinct analysis may be more reliable than ecological regression.”). However, there are problems with both methods. See, e.g., Greiner, *supra*, at 123–39 (arguing that homogeneous precinct analysis is fundamentally flawed and that courts need to be more discerning in dealing with ecological regression).

Redistricting plans based predominantly on racial considerations are unconstitutional unless narrowly tailored to meet a compelling state interest. *Shaw v. Reno*, 509 U.S. 630 (1993). Whether compliance with the Voting Rights Act can be considered a compelling interest is an open ques-

Figure 12. Turnout rate for the white candidate plotted against the percentage of registrants who are white. Precinct-level data, 1982 Democratic Primary for Auditor, Lee County, South Carolina.



Source: Data from James W. Loewen & Bernard Grofman, *Recent Developments in Methods Used in Vote Dilution Litigation*, 21 Urb. Law. 589, 591 tbl.1 (1989).

### D. Statistical Models

Statistical models are widely used in the social sciences and in litigation. For example, the census suffers an undercount, more severe in certain places than others. If some statistical models are to be believed, the undercount can be corrected—moving seats in Congress and millions of dollars a year in tax funds.<sup>134</sup> Other models purport to lift the veil of secrecy from the ballot box, enabling the experts to determine how minority groups have voted—a crucial step in voting rights litigation (*supra* Section V.C). This section discusses the statistical logic of regression models.

A regression model attempts to combine the values of certain variables (the independent variables) to get expected values for another variable (the dependent variable). The model can be expressed in the form of a regression equation. A simple regression equation has only one independent variable; a multiple regression equation has several independent variables. Coefficients in the equation will be interpreted as showing the effects of changing the corresponding variables. This is justified in some situations, as the next example demonstrates.

tion, but efforts to sustain racially motivated redistricting on this basis have not fared well before the Supreme Court. See *Abrams v. Johnson*, 521 U.S. 74 (1997); *Shaw v. Hunt*, 517 U.S. 899 (1996); *Bush v. Vera*, 517 U.S. 952 (1996).

134. See Brown et al., *supra* note 29; *supra* note 89.

Hooke's law (named after Robert Hooke, England, 1653–1703) describes how a spring stretches in response to a load: Strain is proportional to stress. To verify Hooke's law experimentally, a physicist will make a number of observations on a spring. For each observation, the physicist hangs a weight on the spring and measures its length. A statistician could develop a regression model for these data:

$$\text{length} = a + b \times \text{weight} + \epsilon. \quad (1)$$

The error term, denoted by the Greek letter epsilon  $\epsilon$ , is needed because measured length will not be exactly equal to  $a + b \times \text{weight}$ . If nothing else, measurement error must be reckoned with. The model takes  $\epsilon$  as “random error”—behaving like draws made at random with replacement from a box of tickets. Each ticket shows a potential error, which will be realized if that ticket is drawn. The average of the potential errors in the box is assumed to be zero.

Equation (1) has two parameters,  $a$  and  $b$ . These constants of nature characterize the behavior of the spring:  $a$  is length under no load, and  $b$  is elasticity (the increase in length per unit increase in weight). By way of numerical illustration, suppose  $a$  is 400 and  $b$  is 0.05. If the weight is 1, the length of the spring is expected to be

$$400 + 0.05 = 400.05.$$

If the weight is 3, the expected length is

$$400 + 3 \times 0.05 = 400 + 0.15 = 400.15.$$

In either case, the actual length will differ from expected, by a random error  $\epsilon$ .

In standard statistical terminology, the  $\epsilon$ 's for different observations on the spring are assumed to be independent and identically distributed, with a mean of zero. Take the  $\epsilon$ 's for the first two observations. Independence means that the chances for the second  $\epsilon$  do not depend on outcomes for the first. If the errors are like draws made at random with replacement from a box of tickets, as we assumed earlier, that box will not change from one draw to the next—independence. “Identically distributed” means that the chance behavior of the two  $\epsilon$ 's is the same: They are drawn at random from the same box. (See *infra* Appendix for additional discussion.)

The parameters  $a$  and  $b$  in equation (1) are not directly observable, but they can be estimated by the method of least squares.<sup>135</sup> Statisticians often denote esti-

135. It might seem that  $a$  is observable; after all, we can measure the length of the spring with no load. However, the measurement is subject to error, so we observe not  $a$ , but  $a + \epsilon$ . See equation (1). The parameters  $a$  and  $b$  can be estimated, even estimated very well, but they cannot be observed directly. The least squares estimates of  $a$  and  $b$  are the intercept and slope of the regression



mates by hats. Thus,  $\hat{a}$  is the estimate for  $a$ , and  $\hat{b}$  is the estimate for  $b$ . The values of  $\hat{a}$  and  $\hat{b}$  are chosen to minimize the sum of the squared prediction errors. These errors are also called residuals. They measure the difference between the actual length of the spring and the predicted length, the latter being  $\hat{a} + \hat{b} \times \text{weight}$ :

$$\text{actual length} = \hat{a} + \hat{b} \times \text{weight} + \text{residual}. \quad (2)$$

Of course, no one really imagines there to be a box of tickets hidden in the spring. However, the variability of physical measurements (under many but by no means all circumstances) does seem to be remarkably like the variability in draws from a box.<sup>136</sup> In short, the statistical model corresponds rather closely to the empirical phenomenon.

Equation (1) is a statistical model for the data, with unknown parameters  $a$  and  $b$ . The error term  $\epsilon$  is not observable. The model is a theory—and a good one—about how the data are generated. By contrast, equation (2) is a regression equation that is fitted to the data: The intercept  $\hat{a}$ , the slope  $\hat{b}$ , and the residual can all be computed from the data. The results are useful because  $\hat{a}$  is a good estimate for  $a$ , and  $\hat{b}$  is a good estimate for  $b$ . (Similarly, the residual is a good approximation to  $\epsilon$ .) Without the theory, these estimates would be less useful. Is there a theoretical model behind the data processing? Is the model justifiable? These questions can be critical when it comes to making statistical inferences from the data.

In social science applications, statistical models often are invoked without an independent theoretical basis. We give an example involving salary discrimination in the Appendix.<sup>137</sup> The main ideas of such regression modeling can be captured in a hypothetical exchange between a plaintiff seeking to prove salary discrimination and a company denying the allegation. Such a dialog might proceed as follows:

1. Plaintiff argues that the defendant company pays male employees more than females, which establishes a *prima facie* case of discrimination.
2. The company responds that the men are paid more because they are better educated and have more experience.
3. Plaintiff refutes the company's theory by fitting a regression equation that includes a particular, presupposed relationship between salary (the dependent variable) and some measures of education and experience. Plaintiff's expert reports that even after adjusting for differences in education and

line. See *supra* Section V.C.1; Freedman et al., *supra* note 12, at 208–10. The method of least squares was developed by Adrien-Marie Legendre (France, 1752–1833) and Carl Friedrich Gauss (Germany, 1777–1855) to fit astronomical orbits.

136. This is the Gauss model for measurement error. See Freedman et al., *supra* note 12, at 450–52.

137. The Reference Guide to Multiple Regression in this manual describes a comparable example.

experience in this specific manner, men earn more than women. This remaining difference in pay shows discrimination.

4. The company argues that the difference could be the result of chance, not discrimination.
5. Plaintiff replies that because the coefficient for gender in the model is statistically significant, chance is not a good explanation for the data.<sup>138</sup>

In step 3, the three explanatory variables are education (years of schooling completed), experience (years with the firm), and a dummy variable for gender (1 for men and 0 for women). These are supposed to predict salaries (dollars per year). The equation is a formal analog of Hooke's law (equation 1). According to the model, an employee's salary is determined as if by computing

$$a + (b \times \text{education}) + (c \times \text{experience}) + (d \times \text{gender}), \quad (3)$$

and then adding an error  $\varepsilon$  drawn at random from a box of tickets.<sup>139</sup> The parameters  $a$ ,  $b$ ,  $c$ , and  $d$ , are estimated from the data by the method of least squares.

In step 5, the estimated coefficient  $d$  for the dummy variable turns out to be positive and statistically significant and is offered as evidence of disparate impact. Men earn more than women, even after adjusting for differences in background factors that might affect productivity. This showing depends on many assumptions built into the model.<sup>140</sup> Hooke's law—equation (1)—is relatively easy to test experimentally. For the salary discrimination model, validation would be difficult. When expert testimony relies on statistical models, the court may well inquire, what are the assumptions behind the model, and why do they apply to the case at hand? It might then be important to distinguish between two situations:

- The nature of the relationship between the variables is known and regression is being used to make quantitative estimates of parameters in that relationship, or
- The nature of the relationship is largely unknown and regression is being used to determine the nature of the relationship—or indeed whether any relationship exists at all.

138. In some cases, the  $p$ -value has been interpreted as the probability that defendants are innocent of discrimination. However, as noted earlier, such an interpretation is wrong:  $p$  merely represents the probability of getting a large test statistic, given that the model is correct and the true coefficient for gender is zero (see *supra* Section IV.B, *infra* Appendix, Section D.2). Therefore, even if we grant the model, a  $p$ -value less than 50% does not demonstrate a preponderance of the evidence against the null hypothesis.

139. Expression (3) is the expected value for salary, given the explanatory variables (education, experience, gender). The error term is needed to account for deviations from expected: Salaries are not going to be predicted very well by linear combinations of variables such as education and experience.

140. See *infra* Appendix.

Regression was developed to handle situations of the first type, with Hooke's law being an example. The basis for the second type of application is analogical, and the tightness of the analogy is an issue worth exploration.

In employment discrimination cases, and other contexts too, a wide variety of models can be used. This is only to be expected, because the science does not dictate specific equations. In a strongly contested case, each side will have its own model, presented by its own expert. The experts will reach opposite conclusions about discrimination. The dialog might continue with an exchange about which model is better. Although statistical assumptions are challenged in court from time to time, arguments more commonly revolve around the choice of variables. One model may be questioned because it omits variables that should be included—for example, skill levels or prior evaluations.<sup>141</sup> Another model may be challenged because it includes tainted variables reflecting past discriminatory behavior by the firm.<sup>142</sup> The court must decide which model—if either—fits the occasion.<sup>143</sup>

The frequency with which regression models are used is no guarantee that they are the best choice for any particular problem. Indeed, from one perspective, a regression or other statistical model may seem to be a marvel of mathematical rigor. From another perspective, the model is a set of assumptions, supported only by the say-so of the testifying expert. Intermediate judgments are also possible.<sup>144</sup>

141. *E.g.*, *Bazemore v. Friday*, 478 U.S. 385 (1986); *In re Linerboard Antitrust Litig.*, 497 F. Supp. 2d 666 (E.D. Pa. 2007).

142. *E.g.*, *McLaurin v. Nat'l R.R. Passenger Corp.*, 311 F. Supp. 2d 61, 65–66 (D.D.C. 2004) (holding that the inclusion of two allegedly tainted variables was reasonable in light of an earlier consent decree).

143. *E.g.*, *Chang v. Univ. of R.I.*, 606 F. Supp. 1161, 1207 (D.R.I. 1985) (“it is plain to the court that [defendant's] model comprises a better, more useful, more reliable tool than [plaintiff's] counterpart.”); *Presseisen v. Swarthmore College*, 442 F. Supp. 593, 619 (E.D. Pa. 1977) (“[E]ach side has done a superior job in challenging the other's regression analysis, but only a mediocre job in supporting their own . . . and the Court is . . . left with nothing.”), *aff'd*, 582 F.2d 1275 (3d Cir. 1978).

144. *See, e.g.*, David W. Peterson, *Reference Guide on Multiple Regression*, 36 *Jurimetrics J.* 213, 214–15 (1996) (review essay); *see supra* note 21 for references to a range of academic opinion. More recently, some investigators have turned to graphical models. However, these models have serious weaknesses of their own. *See, e.g.*, David A. Freedman, *On Specifying Graphical Models for Causation, and the Identification Problem*, 26 *Evaluation Rev.* 267 (2004).

# Appendix

## A. Frequentists and Bayesians

The mathematical theory of probability consists of theorems derived from axioms and definitions. Mathematical reasoning is seldom controversial, but there may be disagreement as to how the theory should be applied. For example, statisticians may differ on the interpretation of data in specific applications. Moreover, there are two main schools of thought about the foundations of statistics: frequentist and Bayesian (also called objectivist and subjectivist).<sup>145</sup>

Frequentists see probabilities as empirical facts. When a fair coin is tossed, the probability of heads is 1/2; if the experiment is repeated a large number of times, the coin will land heads about one-half the time. If a fair die is rolled, the probability of getting an ace (one spot) is 1/6. If the die is rolled many times, an ace will turn up about one-sixth of the time.<sup>146</sup> Generally, if a chance experiment can be repeated, the relative frequency of an event approaches (in the long run) its probability. By contrast, a Bayesian considers probabilities as representing not facts but degrees of belief: In whole or in part, probabilities are subjective.

Statisticians of both schools use conditional probability—that is, the probability of one event given that another has occurred. For example, suppose a coin is tossed twice. One event is that the coin will land HH. Another event is that at least one H will be seen. Before the coin is tossed, there are four possible, equally likely, outcomes: HH, HT, TH, TT. So the probability of HH is 1/4. However, if we know that at least one head has been obtained, then we can rule out two tails TT. In other words, given that at least one H has been obtained, the conditional probability of TT is 0, and the first three outcomes have conditional probability 1/3 each. In particular, the conditional probability of HH is 1/3. This is usually written as  $P(HH | \text{at least one H}) = 1/3$ . More generally, the probability of an event C is denoted  $P(C)$ ; the conditional probability of D given C is written as  $P(D|C)$ .

Two events C and D are independent if the conditional probability of D given that C occurs is equal to the conditional probability of D given that C does not occur. Statisticians use “ $\sim C$ ” to denote the event that C does not occur. Thus C and D are independent if  $P(D|C) = P(D|\sim C)$ . If C and D are independent, then the probability that both occur is equal to the product of the probabilities:

$$P(C \text{ and } D) = P(C) \times P(D). \quad (A1)$$

145. But see *supra* note 123 (on “objective Bayesianism”).

146. Probabilities may be estimated from relative frequencies, but probability itself is a subtler idea. For example, suppose a computer prints out a sequence of 10 letters H and T (for heads and tails), which alternate between the two possibilities H and T as follows: H T H T H T H T H T. The relative frequency of heads is 5/10 or 50%, but it is not at all obvious that the chance of an H at the next position is 50%. There are difficulties in both the subjectivist and objectivist positions. See Freedman, *supra* note 84.

This is the multiplication rule (or product rule) for independent events. If events are dependent, then conditional probabilities must be used:

$$P(C \text{ and } D) = P(C) \times P(D|C). \quad (\text{A2})$$

This is the multiplication rule for dependent events.

Bayesian statisticians assign probabilities to hypotheses as well as to events; indeed, for them, the distinction between hypotheses and events may not be a sharp one. We turn now to Bayes' rule. If  $H_0$  and  $H_1$  are two hypotheses<sup>147</sup> that govern the probability of an event  $A$ , a Bayesian can use the multiplication rule (A2) to find that

$$P(A \text{ and } H_0) = P(A|H_0)P(H_0) \quad (\text{A3})$$

and

$$P(A \text{ and } H_1) = P(A|H_1)P(H_1). \quad (\text{A4})$$

Moreover,

$$P(A) = P(A \text{ and } H_0) + P(A \text{ and } H_1). \quad (\text{A5})$$

The multiplication rule (A2) also shows that

$$P(H_1|A) = \frac{P(A \text{ and } H_1)}{P(A)}. \quad (\text{A6})$$

We use (A4) to evaluate  $P(A \text{ and } H_1)$  in the numerator of (A6), and (A3), (A4), and (A5) to evaluate  $P(A)$  in the denominator:

$$P(H_1|A) = \frac{P(A|H_1)P(H_1)}{P(A|H_0)P(H_0) + P(A|H_1)P(H_1)}. \quad (\text{A7})$$

This is a special case of Bayes' rule. It yields the conditional probability of hypothesis  $H_0$  given that event  $A$  has occurred.

For a stylized example in a criminal case,  $H_0$  is the hypothesis that blood found at the scene of a crime came from a person other than the defendant;  $H_1$  is the hypothesis that the blood came from the defendant;  $A$  is the event that blood from the crime scene and blood from the defendant are both type A. Then  $P(H_0)$  is the prior probability of  $H_0$ , based on subjective judgment, while  $P(H_0|A)$  is the posterior probability—updated from the prior using the data.

147.  $H_0$  is read "H-sub-zero," while  $H_1$  is "H-sub-one."

Type A blood occurs in 42% of the population. So  $P(A|H_0) = 0.42$ .<sup>148</sup> Because the defendant has type A blood,  $P(A|H_1) = 1$ . Suppose the prior probabilities are  $P(H_0) = P(H_1) = 0.5$ . According to (A7), the posterior probability that the blood is from the defendant is

$$P(H_1|A) = \frac{1 \times 0.5}{0.42 \times 0.5 + 1 \times 0.5} = 0.70. \quad (\text{A8})$$

Thus, the data increase the likelihood that the blood is the defendant's. The probability went up from the prior value of  $P(H_1) = 0.50$  to the posterior value of  $P(H_1|A) = 0.70$ .

More generally,  $H_0$  and  $H_1$  refer to parameters in a statistical model. For a stylized example in an employment discrimination case,  $H_0$  asserts equal selection rates in a population of male and female applicants;  $H_1$  asserts that the selection rates are not equal;  $A$  is the event that a test statistic exceeds 2 in absolute value. In such situations, the Bayesian proceeds much as before. However, the frequentist computes  $P(A|H_0)$ , and rejects  $H_0$  if this probability falls below 5%. Frequentists have to stop there, because they view  $P(H_0|A)$  as poorly defined at best. In their setup,  $P(H_0)$  and  $P(H_1)$  rarely make sense, and these prior probabilities are needed to compute  $P(H_1|A)$ : See *supra* equation (A7).

Assessing probabilities, conditional probabilities, and independence is not entirely straightforward, either for frequentists or Bayesians. Inquiry into the basis for expert judgment may be useful, and casual assumptions about independence should be questioned.<sup>149</sup>

## B. The Spock Jury: Technical Details

The rest of this Appendix provides some technical backup for the examples in Sections IV and V, *supra*. We begin with the *Spock* jury case. On the null hypothesis, a sample of 350 people was drawn at random from a large population that was 50% male and 50% female. The number of women in the sample follows the binomial distribution. For example, the chance of getting exactly 102 women in the sample is given by the binomial formula<sup>150</sup>

$$\frac{n!}{j! \times (n-j)!} f^j (1-f)^{n-j}. \quad (\text{A9})$$

148. Not all statisticians would accept the identification of a population frequency with  $P(A|H_0)$ . Indeed,  $H_0$  has been translated into a hypothesis that the true donor has been selected from the population at random (i.e., in a manner that is uncorrelated with blood type). This step needs justification. See *supra* note 123.

149. For problematic assumptions of independence in litigation, see, e.g., *Wilson v. State*, 803 A.2d 1034 (Md. 2002) (error to admit multiplied probabilities in a case involving two deaths of infants in same family); 1 McCormick, *supra* note 2, § 210; see also *supra* note 29 (on census litigation).

150. The binomial formula is discussed in, e.g., Freedman et al., *supra* note 12, at 255–61.

In the formula,  $n$  stands for the sample size, and so  $n = 350$ ; and  $j = 102$ . The  $f$  is the fraction of women in the population; thus,  $f = 0.50$ . The exclamation point denotes factorials:  $1! = 1$ ,  $2! = 2 \times 1 = 2$ ,  $3! = 3 \times 2 \times 1 = 6$ , and so forth. The chance of 102 women works out to  $10^{-15}$ . In the same way, we can compute the chance of getting 101 women, or 100, or any other particular number. The chance of getting 102 women or fewer is then computed by addition. The chance is  $p = 2 \times 10^{-15}$ , as reported *supra* note 98. This is very bad news for the null hypothesis.

With the binomial distribution given by (9), the expected the number of women in the sample is

$$nf = 350 \times 0.5 = 175. \quad (\text{A10})$$

The standard error is

$$\sqrt{n} \times \sqrt{f \times (1-f)} = \sqrt{350} \times \sqrt{0.5 \times 0.5} = 9.35. \quad (\text{A11})$$

The observed value of 102 is nearly 8 SEs below the expected value, which is a lot of SEs.

Figure 13 shows the probability histogram for the number of women in the sample.<sup>151</sup> The graph is drawn so that the area between two values is proportional to the chance that the number of women will fall in that range. For example, take the rectangle over 175; its base covers the interval from 174.5 to 175.5. The area of this rectangle is 4.26% of the total area. So the chance of getting exactly 175 women is 4.26%. Next, take the range from 165 to 185 (inclusive): 73.84% of the area falls into this range. This means there is a 73.84% chance that the number of women in the sample will be in the range from 165 to 185 (inclusive).

According to a fundamental theorem in statistics (the central limit theorem), the histogram follows the normal curve.<sup>152</sup> Figure 13 shows the curve for comparison: The normal curve is almost indistinguishable from the top of the histogram. For a numerical example, suppose the jury panel had included 155 women. On the null hypothesis, there is about a 1.85% chance of getting 155 women or fewer. The normal curve gives 1.86%. The error is nil. Ordinarily, we would just report  $p = 2\%$ , as in the text (*supra* Section IV.B.1).

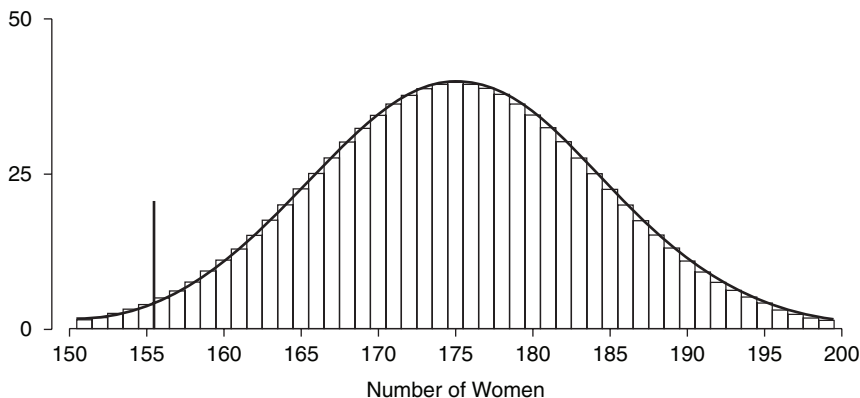
Finally, we consider power. Suppose we reject the null hypothesis when the number of women in the sample is 155 or less. Let us assume a particular alternative hypothesis that quantifies the degree of discrimination against women: The jury panel is selected at random from a population that is 40% female, rather than 50%. Figure 14 shows the probability histogram for the number of women, but now the histogram is computed according to the alternative hypothesis. Again,

151. Probability histograms are discussed in, e.g., *id.* at 310–13.

152. The central limit theorem is discussed in, e.g., *id.* at 315–27.

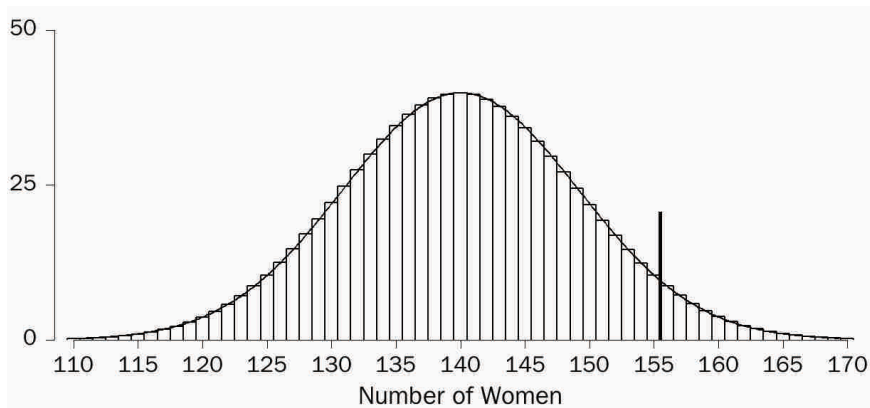
Reference Guide on Statistics

Figure 13. Probability histogram for the number of women in a random sample of 350 people drawn from a large population that is 50% female and 50% male. The normal curve is shown for comparison. About 2% of the area under the histogram is to the left of 155 (marked by a heavy vertical line).



*Note:* The vertical line is placed at 155.5, and so the area to the left of it includes the rectangles over 155, 154, . . . ; the area represents the chance of getting 155 women or fewer. *Cf.* Freedman et al., *supra* note 12, at 317. The units on the vertical axis are “percent per standard unit”; *cf. id.* at 80, 315.

Figure 14. Probability histogram for the number of women in a random sample of 350 people drawn from a large population that is 40% female and 60% male. The normal curve is shown for comparison. The area to the left of 155 (marked by a heavy vertical line) is about 95%.





the histogram follows the normal curve. About 95% of the area is to the left of 155, and so power is about 95%. The area can be computed exactly by using the binomial distribution, or to an excellent approximation using the normal curve.

Figures 13 and 14 have the same shape: The central limit theorem is at work. However, the histograms are centered differently. Figure 13 is centered at 175, according to requirements of the null hypothesis. Figure 14 is centered at 140, because the alternative hypothesis is used to determine the center, not the null hypothesis. Thus, 155 is well to the left of center in Figure 13, and well to the right in Figure 14: The figures have different centers. The main point of Figures 13 and 14 is that chances can often be approximated by areas under the normal curve, justifying the large-sample theory presented *supra* Sections IV.A–B.

### C. The Nixon Papers: Technical Details

With the Nixon papers, the population consists of 20,000 boxes. A random sample of 500 boxes is drawn and each sample box is appraised. Statistical theory enables us to make some precise statements about the behavior of the sample average.

- The expected value of the sample average equals the population average. Even more tersely, the sample average is an unbiased estimate of the population average.
- The standard error for the sample average equals

$$\sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}. \quad (\text{A12})$$

In (A12), the  $N$  stands for the size of the population, which is 20,000; and  $n$  stands for the size of the sample, which is 500. The first factor in (A12), with the square root, is the finite sample correction factor. Here, as in many other such examples, the correction factor is so close to 1 that it can safely be ignored. (This is why the size of population usually has no bearing on the precision of the sample average as an estimator for the population average.) Next,  $\sigma$  is the population standard deviation. This is unknown, but it can be estimated by the sample standard deviation, which is \$2200. The SE for the sample mean is therefore estimated from the data as  $\$2200/\sqrt{500}$ , which is nearly \$100. Plaintiff's total claim is 20,000 times the sample average. The SE for the total claim is therefore  $20,000 \times \$100 = \$2,000,000$ . (Here, the size of the population comes into the formula.)

With a large sample, the probability histogram for the sample average follows the normal curve quite closely. That is a consequence of the central limit theorem. The center of the histogram is the population average. The SE is given by (A12), and is about \$100.

- What is the chance that the sample average differs from the population average by 1 SE or less? This chance is equal to the area under the probability histogram within 1 SE of average, which by the central limit theorem is almost equal to the area under the standard normal curve between  $-1$  and  $1$ ; that normal area is about 68%.
- What is the chance that the sample average differs from the population average by 2 SE or less? By the same reasoning, this chance is about equal to the area under the standard normal curve between  $-2$  and  $2$ , which is about 95%.
- What is the chance that the sample average differs from the population average by 3 SE or less? This chance is about equal to the area under the standard normal curve between  $-3$  and  $3$ , which is about 99.7%.

To sum up, the probability histogram for the sample average is centered at the population average. The spread is given by the standard error. The histogram follows the normal curve. That is why confidence levels can be based on the standard error, with confidence levels read off the normal curve—for estimators that are essentially unbiased, and obey the central limit theorem (*supra* Section IV.A.2, Appendix Section B).<sup>153</sup> These large-sample methods generally work for sums, averages, and rates, although much depends on the design of the sample.

More technically, the normal curve is the density of a normal distribution. The standard normal density has mean equal to 0 and standard error equal to 1. Its equation is

$$y = e^{-x^2/2} / \sqrt{2\pi}$$

where  $e = 2.71828\dots$  and  $\pi = 3.14159\dots$ . This density can be rescaled to have any desired mean and standard error. The resulting densities are the famous “normal curves” or “bell-shaped curves” of statistical theory. In Figure 12, the density is scaled to match the probability histogram in terms of the mean and standard error; likewise in Figure 13.

## D. A Social Science Example of Regression: Gender Discrimination in Salaries

### 1. The regression model

To illustrate social science applications of the kind that might be seen in litigation, Section V referred to a stylized example on salary discrimination. A particular

153. See, e.g., *id.* at 409–24. On the standard deviation, see *supra* Section III.E; Freedman et al., *supra* note 12, at 67–72. The finite sample correction factor is discussed in *id.* at 367–70.

regression model was used to predict salaries (dollars per year) of employees in a firm. It had three explanatory variables: education (years of schooling completed), experience (years with the firm), and a dummy variable for gender (1 for men and 0 for women). The regression equation is

$$\text{salary} = a + b \times \text{education} + c \times \text{experience} + d \times \text{gender} + \varepsilon. \quad (\text{A13})$$

Equation (A13) is a statistical model for the data, with unknown parameters  $a$ ,  $b$ ,  $c$ , and  $d$ . Here,  $a$  is the intercept and the other parameters are regression coefficients. The  $\varepsilon$  at the end of the equation is an unobservable error term. In the right-hand side of (A3) and similar expressions, by convention, the multiplications are done before the additions.

As noted in Section V, the equation is a formal analog of Hooke's law (1). According to the model, an employee's salary is determined as if by computing

$$a + b \times \text{education} + c \times \text{experience} + d \times \text{gender} \quad (\text{A14})$$

and then adding an error  $\varepsilon$  drawn at random from a box of tickets. Expression (A14) is the expected value for salary, given the explanatory variables (education, experience, gender). The error term is needed to account for deviations from expected: Salaries are not going to be predicted very well by linear combinations of variables such as education and experience.

The parameters are estimated from the data using least squares. If the estimated coefficient for the dummy variable turns out to be positive and statistically significant, that would be evidence of disparate impact. Men earn more than women, even after adjusting for differences in background factors that might affect productivity. Suppose the estimated equation turns out as follows:

$$\begin{aligned} \text{predicted salary} = & \$7100 + \$1300 \times \text{education} + \$2200 \\ & \times \text{experience} + \$700 \times \text{gender}. \end{aligned} \quad (\text{A15})$$

According to (A15), the estimated value for the intercept  $a$  in (A14) is \$7100; the estimated value for the coefficient  $b$  is \$1300, and so forth. According to equation (A15), every extra year of education is worth \$1300. Similarly, every extra year of experience is worth \$2200. And, most important, the company gives men a salary premium of \$700 over women with the same education and experience.

A male employee with 12 years of education (high school) and 10 years of experience, for example, would have a predicted salary of

$$\begin{aligned} & \$7100 + \$1300 \times 12 + \$2200 \times 10 + \$700 \times 1 \\ & = \$7100 + \$15,600 + \$22,000 + \$700 = \$45,400. \end{aligned} \quad (\text{A16})$$

A similarly situated female employee has a predicted salary of only

$$\begin{aligned}
 & \$7100 + \$1300 \times 12 + \$2200 \times 10 + \$700 \times 0 \\
 & = \$7100 + \$15,600 + \$22,000 + \$0 = \$44,700.
 \end{aligned}
 \tag{A17}$$

Notice the impact of the gender variable in the model: \$700 is added to equation (A16), but not to equation (A17).

A major step in proving discrimination is showing that the estimated coefficient of the gender variable—\$700 in the numerical illustration—is statistically significant. This showing depends on the assumptions built into the model. Thus, each extra year of education is assumed to be worth the same across all levels of experience. Similarly, each extra year of experience is worth the same across all levels of education. Furthermore, the premium paid to men does not depend systematically on education or experience. Omitted variables such as ability, quality of education, or quality of experience do not make any systematic difference to the predictions of the model.<sup>154</sup> These are all assumptions made going into the analysis, rather than conclusions coming out of the data.

Assumptions are also made about the error term—the mysterious  $\epsilon$  at the end of (A13). The errors are assumed to be independent and identically distributed from person to person in the dataset. Such assumptions are critical when computing  $p$ -values and demonstrating statistical significance. Regression modeling that does not produce statistically significant coefficients will not be good evidence of discrimination, and statistical significance cannot be established unless stylized assumptions are made about unobservable error terms.

The typical regression model, like the one sketched above, therefore involves a host of assumptions. As noted in Section V, Hooke's law—equation (1)—is relatively easy to test experimentally. For the salary discrimination model—equation (A13)—validation would be difficult. That is why we suggested that when expert testimony relies on statistical models, the court may well inquire about the assumptions behind the model and why they apply to the case at hand.

## 2. Standard errors, $t$ -statistics, and statistical significance

Statistical proof of discrimination depends on the significance of the estimated coefficient for the gender variable. Significance is determined by the  $t$ -test, using the standard error. The standard error measures the likely difference between the estimated value for the coefficient and its true value. The estimated value is \$700—the coefficient of the gender variable in equation (A5); the true value  $d$  in (A13), remains unknown. According to the model, the difference between the estimated value and the true value is due to the action of the error term  $\epsilon$  in (A3). Without  $\epsilon$ , observed values would line up perfectly with expected values,

154. Technically, these omitted variables are assumed to be independent of the error term in the equation.

and estimated values for parameters would be exactly equal to true values. This does not happen.

The  $t$ -statistic is the estimated value divided by its standard error. For example, in (A15), the estimate for  $d$  is \$700. If the standard error is \$325, then  $t$  is  $\$700/\$325 = 2.15$ . This is significant—that is, hard to explain as the product of random error. Under the null hypothesis that  $d$  is zero, there is only about a 5% chance that the absolute value of  $t$  is greater than 2. (We are assuming the sample is large.) Thus, statistical significance is achieved (*supra* Section IV.B.2). Significance would be taken as evidence that  $d$ —the true parameter in the model (A13)—does not vanish. According to a viewpoint often presented in the social science journals and the courtroom, here is statistical proof that gender matters in determining salaries. On the other hand, if the standard error is \$1400, then  $t$  is  $\$700/\$1400 = 0.5$ . The difference between the estimated value of  $d$  and zero could easily result from chance. So the true value of  $d$  could well be zero, in which case gender does not affect salaries.

Of course, the parameter  $d$  is only a construct in a model. If the model is wrong, the standard error,  $t$ -statistic, and significance level are rather difficult to interpret. Even if the model is granted, there is a further issue. The 5% is the chance that the absolute value of  $t$  exceeds 2, given the model and given the null hypothesis that  $d$  is zero. However, the 5% is often taken to be the chance of the null hypothesis given the data. This misinterpretation is commonplace in the social science literature, and it appears in some opinions describing expert testimony.<sup>155</sup> For a frequentist statistician, the chance that  $d$  is zero given the data makes no sense: Parameters do not exhibit chance variation. For a Bayesian statistician, the chance that  $d$  is zero given the data makes good sense, but the computation via the  $t$ -test could be seriously in error, because the prior probability that  $d$  is zero has not been taken into account.<sup>156</sup>

The mathematical terminology in the previous paragraph may need to be deciphered: The “absolute value” of  $t$  is the magnitude, ignoring sign. Thus, the absolute value of both +3 and −3 is 3.

155. See *supra* Section IV.B & notes 102 & 116.

156. See *supra* Section IV & *supra* Appendix.

## Glossary of Terms

The following definitions are adapted from a variety of sources, including Michael O. Finkelstein & Bruce Levin, *Statistics for Lawyers* (2d ed. 2001), and David A. Freedman et al., *Statistics* (4th ed. 2007).

**absolute value.** Size, neglecting sign. The absolute value of +2.7 is 2.7; so is the absolute value of −2.7.

**adjust for.** See control for.

**alpha ( $\alpha$ ).** A symbol often used to denote the probability of a Type I error. See Type I error; size. Compare beta.

**alternative hypothesis.** A statistical hypothesis that is contrasted with the null hypothesis in a significance test. See statistical hypothesis; significance test.

**area sample.** A probability sample in which the sampling frame is a list of geographical areas. That is, the researchers make a list of areas, choose some at random, and interview people in the selected areas. This is a cost-effective way to draw a sample of people. See probability sample; sampling frame.

**arithmetic mean.** See mean.

**average.** See mean.

**Bayes' rule.** In its simplest form, an equation involving conditional probabilities that relates a “prior probability” known or estimated before collecting certain data to a “posterior probability” that reflects the impact of the data on the prior probability. In Bayesian statistical inference, “the prior” expresses degrees of belief about various hypotheses. Data are collected according to some statistical model; at least, the model represents the investigator’s beliefs. Bayes’ rule combines the prior with the data to yield the posterior probability, which expresses the investigator’s beliefs about the parameters, given the data. See Appendix A. Compare frequentist.

**beta ( $\beta$ ).** A symbol sometimes used to denote power, and sometimes to denote the probability of a Type II error. See Type II error; power. Compare alpha.

**between-observer variability.** Differences that occur when two observers measure the same thing. Compare within-observer variability.

**bias.** Also called systematic error. A systematic tendency for an estimate to be too high or too low. An estimate is unbiased if the bias is zero. (Bias does not mean prejudice, partiality, or discriminatory intent.) See nonsampling error. Compare sampling error.

**bin.** A class interval in a histogram. See class interval; histogram.

**binary variable.** A variable that has only two possible values (e.g., gender). Called a dummy variable when the two possible values are 0 and 1.

**binomial distribution.** A distribution for the number of occurrences in repeated, independent “trials” where the probabilities are fixed. For example, the num-

ber of heads in 100 tosses of a coin follows a binomial distribution. When the probability is not too close to 0 or 1 and the number of trials is large, the binomial distribution has about the same shape as the normal distribution. See normal distribution; Poisson distribution.

**blind.** See double-blind experiment.

**bootstrap.** Also called resampling; Monte Carlo method. A procedure for estimating sampling error by constructing a simulated population on the basis of the sample, then repeatedly drawing samples from the simulated population.

**categorical data; categorical variable.** See qualitative variable. Compare quantitative variable.

**central limit theorem.** Shows that under suitable conditions, the probability histogram for a sum (or average or rate) will follow the normal curve. See histogram; normal curve.

**chance error.** See random error; sampling error.

**chi-squared ( $\chi^2$ ).** The chi-squared statistic measures the distance between the data and expected values computed from a statistical model. If the chi-squared statistic is too large to explain by chance, the data contradict the model. The definition of “large” depends on the context. See statistical hypothesis; significance test.

**class interval.** Also, bin. The base of a rectangle in a histogram; the area of the rectangle shows the percentage of observations in the class interval. See histogram.

**cluster sample.** A type of random sample. For example, investigators might take households at random, then interview all people in the selected households. This is a cluster sample of people: A cluster consists of all the people in a selected household. Generally, clustering reduces the cost of interviewing. See multistage cluster sample.

**coefficient of determination.** A statistic (more commonly known as *R*-squared) that describes how well a regression equation fits the data. See *R*-squared.

**coefficient of variation.** A statistic that measures spread relative to the mean: SD/mean, or SE/expected value. See expected value; mean; standard deviation; standard error.

**collinearity.** See multicollinearity.

**conditional probability.** The probability that one event will occur given that another has occurred.

**confidence coefficient.** See confidence interval.

**confidence interval.** An estimate, expressed as a range, for a parameter. For estimates such as averages or rates computed from large samples, a 95% confidence interval is the range from about two standard errors below to two standard errors above the estimate. Intervals obtained this way cover the true

value about 95% of the time, and 95% is the confidence level or the confidence coefficient. See central limit theorem; standard error.

**confidence level.** See confidence interval.

**confounding variable; confounder.** A confounder is correlated with the independent variable and the dependent variable. An association between the dependent and independent variables in an observational study may not be causal, but may instead be due to confounding. See controlled experiment; observational study.

**consistent estimator.** An estimator that tends to become more and more accurate as the sample size grows. Inconsistent estimators, which do not become more accurate as the sample gets larger, are frowned upon by statisticians.

**content validity.** The extent to which a skills test is appropriate to its intended purpose, as evidenced by a set of questions that adequately reflect the domain being tested. See validity. Compare reliability.

**continuous variable.** A variable that has arbitrarily fine gradations, such as a person's height. Compare discrete variable.

**control for.** Statisticians may control for the effects of confounding variables in nonexperimental data by making comparisons for smaller and more homogeneous groups of subjects, or by entering the confounders as explanatory variables in a regression model. To “adjust for” is perhaps a better phrase in the regression context, because in an observational study the confounding factors are not under experimental control; statistical adjustments are an imperfect substitute. See regression model.

**control group.** See controlled experiment.

**controlled experiment.** An experiment in which the investigators determine which subjects are put into the treatment group and which are put into the control group. Subjects in the treatment group are exposed by the investigators to some influence—the treatment; those in the control group are not so exposed. For example, in an experiment to evaluate a new drug, subjects in the treatment group are given the drug, and subjects in the control group are given some other therapy; the outcomes in the two groups are compared to see whether the new drug works.

Randomization—that is, randomly assigning subjects to each group—is usually the best way to ensure that any observed difference between the two groups comes from the treatment rather than from preexisting differences. Of course, in many situations, a randomized controlled experiment is impractical, and investigators must then rely on observational studies. Compare observational study.

**convenience sample.** A nonrandom sample of units, also called a grab sample. Such samples are easy to take but may suffer from serious bias. Typically, mall samples are convenience samples.



**correlation coefficient.** A number between  $-1$  and  $1$  that indicates the extent of the linear association between two variables. Often, the correlation coefficient is abbreviated as  $r$ .

**covariance.** A quantity that describes the statistical interrelationship of two variables. Compare correlation coefficient; standard error; variance.

**covariate.** A variable that is related to other variables of primary interest in a study; a measured confounder; a statistical control in a regression equation.

**criterion.** The variable against which an examination or other selection procedure is validated. See validity.

**data.** Observations or measurements, usually of units in a sample taken from a larger population.

**degrees of freedom.** See  $t$ -test.

**dependence.** Two events are dependent when the probability of one is affected by the occurrence or non-occurrence of the other. Compare independence; dependent variable.

**dependent variable.** Also called outcome variable. Compare independent variable.

**descriptive statistics.** Like the mean or standard deviation, used to summarize data.

**differential validity.** Differences in validity across different groups of subjects. See validity.

**discrete variable.** A variable that has only a small number of possible values, such as the number of automobiles owned by a household. Compare continuous variable.

**distribution.** See frequency distribution; probability distribution; sampling distribution.

**disturbance term.** A synonym for error term.

**double-blind experiment.** An experiment with human subjects in which neither the diagnosticians nor the subjects know who is in the treatment group or the control group. This is accomplished by giving a placebo treatment to patients in the control group. In a single-blind experiment, the patients do not know whether they are in treatment or control; the diagnosticians have this information.

**dummy variable.** Generally, a dummy variable takes only the values  $0$  or  $1$ , and distinguishes one group of interest from another. See binary variable; regression model.

**econometrics.** Statistical study of economic issues.

**epidemiology.** Statistical study of disease or injury in human populations.

**error term.** The part of a statistical model that describes random error, i.e., the impact of chance factors unrelated to variables in the model. In econometrics, the error term is called a disturbance term.

**estimator.** A sample statistic used to estimate the value of a population parameter. For example, the sample average commonly is used to estimate the population average. The term “estimator” connotes a statistical procedure, whereas an “estimate” connotes a particular numerical result.

**expected value.** See random variable.

**experiment.** See controlled experiment; randomized controlled experiment. Compare observational study.

**explanatory variable.** See independent variable; regression model.

**external validity.** See validity.

**factors.** See independent variable.

**Fisher’s exact test.** A statistical test for comparing two sample proportions. For example, take the proportions of white and black employees getting a promotion. An investigator may wish to test the null hypothesis that promotion does not depend on race. Fisher’s exact test is one way to arrive at a  $p$ -value. The calculation is based on the hypergeometric distribution. For details, see Michael O. Finkelstein and Bruce Levin, *Statistics for Lawyers* 154–56 (2d ed. 2001). See hypergeometric distribution;  $p$ -value; significance test; statistical hypothesis.

**fitted value.** See residual.

**fixed significance level.** Also alpha; size. A preset level, such as 5% or 1%; if the  $p$ -value of a test falls below this level, the result is deemed statistically significant. See significance test. Compare observed significance level;  $p$ -value.

**frequency; relative frequency.** Frequency is the number of times that something occurs; relative frequency is the number of occurrences, relative to a total. For example, if a coin is tossed 1000 times and lands heads 517 times, the frequency of heads is 517; the relative frequency is 0.517, or 51.7%.

**frequency distribution.** Shows how often specified values occur in a dataset.

**frequentist.** Also called objectivist. Describes statisticians who view probabilities as objective properties of a system that can be measured or estimated. Compare Bayesian. See Appendix.

**Gaussian distribution.** A synonym for the normal distribution. See normal distribution.

**general linear model.** Expresses the dependent variable as a linear combination of the independent variables plus an error term whose components may be dependent and have differing variances. See error term; linear combination; variance. Compare regression model.

**grab sample.** See convenience sample.

**heteroscedastic.** See scatter diagram.

**highly significant.** See  $p$ -value; practical significance; significance test.

**histogram.** A plot showing how observed values fall within specified intervals, called bins or class intervals. Generally, matters are arranged so that the area under the histogram, but over a class interval, gives the frequency or relative frequency of data in that interval. With a probability histogram, the area gives the chance of observing a value that falls in the corresponding interval.

**homoscedastic.** See scatter diagram.

**hypergeometric distribution.** Suppose a sample is drawn at random, without replacement, from a finite population. How many times will items of a certain type come into the sample? The hypergeometric distribution gives the probabilities. For more details, see 1 William Feller, *An Introduction to Probability Theory and Its Applications* 41–42 (2d ed. 1957). Compare Fisher’s exact test.

**hypothesis.** See alternative hypothesis; null hypothesis; one-sided hypothesis; significance test; statistical hypothesis; two-sided hypothesis.

**hypothesis test.** See significance test.

**identically distributed.** Random variables are identically distributed when they have the same probability distribution. For example, consider a box of numbered tickets. Draw tickets at random with replacement from the box. The draws will be independent and identically distributed.

**independence.** Also, statistical independence. Events are independent when the probability of one is unaffected by the occurrence or non-occurrence of the other. Compare conditional probability; dependence; independent variable; dependent variable.

**independent variable.** Independent variables (also called explanatory variables, predictors, or risk factors) represent the causes and potential confounders in a statistical study of causation; the dependent variable represents the effect. In an observational study, independent variables may be used to divide the population up into smaller and more homogenous groups (“stratification”). In a regression model, the independent variables are used to predict the dependent variable. For example, the unemployment rate has been used as the independent variable in a model for predicting the crime rate; the unemployment rate is the independent variable in this model, and the crime rate is the dependent variable. The distinction between independent and dependent variables is unrelated to statistical independence. See regression model. Compare dependent variable; dependence; independence.

**indicator variable.** See dummy variable.

**internal validity.** See validity.

**interquartile range.** Difference between 25th and 75th percentile. See percentile.

**interval estimate.** A confidence interval, or an estimate coupled with a standard error. See confidence interval; standard error. Compare point estimate.

**least squares.** See least squares estimator; regression model.

**least squares estimator.** An estimator that is computed by minimizing the sum of the squared residuals. See residual.

**level.** The level of a significance test is denoted alpha ( $\alpha$ ). See alpha; fixed significance level; observed significance level;  $p$ -value; significance test.

**linear combination.** To obtain a linear combination of two variables, multiply the first variable by some constant, multiply the second variable by another constant, and add the two products. For example,  $2u + 3v$  is a linear combination of  $u$  and  $v$ .

**list sample.** See systematic sample.

**loss function.** Statisticians may evaluate estimators according to a mathematical formula involving the errors—that is, differences between actual values and estimated values. The “loss” may be the total of the squared errors, or the total of the absolute errors, etc. Loss functions seldom quantify real losses, but may be useful summary statistics and may prompt the construction of useful statistical procedures. Compare risk.

**lurking variable.** See confounding variable.

**mean.** Also, the average; the expected value of a random variable. The mean gives a way to find the center of a batch of numbers: Add the numbers and divide by how many there are. Weights may be employed, as in “weighted mean” or “weighted average.” See random variable. Compare median; mode.

**measurement validity.** See validity. Compare reliability.

**median.** The median, like the mean, is a way to find the center of a batch of numbers. The median is the 50th percentile. Half the numbers are larger, and half are smaller. (To be very precise: at least half the numbers are greater than or equal to the median; At least half the numbers are less than or equal to the median; for small datasets, the median may not be uniquely defined.) Compare mean; mode; percentile.

**meta-analysis.** Attempts to combine information from all studies on a certain topic. For example, in the epidemiological context, a meta-analysis may attempt to provide a summary odds ratio and confidence interval for the effect of a certain exposure on a certain disease.

**mode.** The most common value. Compare mean; median.

**model.** See probability model; regression model; statistical model.

**multicollinearity.** Also, collinearity. The existence of correlations among the independent variables in a regression model. See independent variable; regression model.

**multiple comparison.** Making several statistical tests on the same dataset. Multiple comparisons complicate the interpretation of a  $p$ -value. For example, if 20 divisions of a company are examined, and one division is found to have a disparity significant at the 5% level, the result is not surprising; indeed, it would be expected under the null hypothesis. Compare  $p$ -value; significance test; statistical hypothesis.

**multiple correlation coefficient.** A number that indicates the extent to which one variable can be predicted as a linear combination of other variables. Its magnitude is the square root of  $R$ -squared. See linear combination;  $R$ -squared; regression model. Compare correlation coefficient.

**multiple regression.** A regression equation that includes two or more independent variables. See regression model. Compare simple regression.

**multistage cluster sample.** A probability sample drawn in stages, usually after stratification; the last stage will involve drawing a cluster. See cluster sample; probability sample; stratified random sample.

**multivariate methods.** Methods for fitting models with multiple variables; in statistics, multiple response variables; in other fields, multiple explanatory variables. See regression model.

**natural experiment.** An observational study in which treatment and control groups have been formed by some natural development; the assignment of subjects to groups is akin to randomization. See observational study. Compare controlled experiment.

**nonresponse bias.** Systematic error created by differences between respondents and nonrespondents. If the nonresponse rate is high, this bias may be severe.

**nonsampling error.** A catch-all term for sources of error in a survey, other than sampling error. Nonsampling errors cause bias. One example is selection bias: The sample is drawn in a way that tends to exclude certain subgroups in the population. A second example is nonresponse bias: People who do not respond to a survey are usually different from respondents. A final example: Response bias arises, for example, if the interviewer uses a loaded question.

**normal distribution.** Also, Gaussian distribution. When the normal distribution has mean equal to 0 and standard error equal to 1, it is said to be “standard normal.” The equation for the density is then

$$y = e^{-x^2/2} / \sqrt{2\pi}$$

where  $e = 2.71828\dots$  and  $\pi = 3.14159\dots$ . The density can be rescaled to have any desired mean and standard error, resulting in the famous “bell-shaped curves” of statistical theory. Terminology notwithstanding, there need be nothing wrong with a distribution that differs from normal.

**null hypothesis.** For example, a hypothesis that there is no difference between two groups from which samples are drawn. See significance test; statistical hypothesis. Compare alternative hypothesis.

**objectivist.** See frequentist.

**observational study.** A study in which subjects select themselves into groups; investigators then compare the outcomes for the different groups. For example, studies of smoking are generally observational. Subjects decide whether or not to smoke; the investigators compare the death rate for smokers to the death rate for nonsmokers. In an observational study, the groups may differ in important ways that the investigators do not notice; controlled experiments minimize this problem. The critical distinction is that in a controlled experiment, the investigators intervene to manipulate the circumstances of the subjects; in an observational study, the investigators are passive observers. (Of course, running a good observational study is hard work, and may be quite useful.) Compare confounding variable; controlled experiment.

**observed significance level.** A synonym for  $p$ -value. See significance test. Compare fixed significance level.

**odds.** The probability that an event will occur divided by the probability that it will not. For example, if the chance of rain tomorrow is  $2/3$ , then the odds on rain are  $(2/3)/(1/3) = 2/1$ , or 2 to 1; the odds against rain are 1 to 2.

**odds ratio.** A measure of association, often used in epidemiology. For example, if 10% of all people exposed to a chemical develop a disease, compared with 5% of people who are not exposed, then the odds of the disease in the exposed group are  $10/90 = 1/9$ , compared with  $5/95 = 1/19$  in the unexposed group. The odds ratio is  $(1/9)/(1/19) = 19/9 = 2.1$ . An odds ratio of 1 indicates no association. Compare relative risk.

**one-sided hypothesis; one-tailed hypothesis.** Excludes the possibility that a parameter could be, for example, less than the value asserted in the null hypothesis. A one-sided hypothesis leads to a one-sided (or one-tailed) test. See significance test; statistical hypothesis; compare two-sided hypothesis.

**one-sided test; one-tailed test.** See one-sided hypothesis.

**outcome variable.** See dependent variable.

**outlier.** An observation that is far removed from the bulk of the data. Outliers may indicate faulty measurements and they may exert undue influence on summary statistics, such as the mean or the correlation coefficient.

**$p$ -value.** Result from a statistical test. The probability of getting, just by chance, a test statistic as large as or larger than the observed value. Large  $p$ -values are consistent with the null hypothesis; small  $p$ -values undermine the null hypothesis. However,  $p$  does not give the probability that the null hypothesis is true. If  $p$  is smaller than 5%, the result is statistically significant. If  $p$  is smaller

than 1%, the result is highly significant. The  $p$ -value is also called the observed significance level. See significance test; statistical hypothesis.

**parameter.** A numerical characteristic of a population or a model. See probability model.

**percentile.** To get the percentiles of a dataset, array the data from the smallest value to the largest. Take the 90th percentile by way of example: 90% of the values fall below the 90th percentile, and 10% are above. (To be very precise: At least 90% of the data are at the 90th percentile or below; at least 10% of the data are at the 90th percentile or above.) The 50th percentile is the median: 50% of the values fall below the median, and 50% are above. On the LSAT, a score of 152 places a test taker at the 50th percentile; a score of 164 is at the 90th percentile; a score of 172 is at the 99th percentile. Compare mean; median; quartile.

**placebo.** See double-blind experiment.

**point estimate.** An estimate of the value of a quantity expressed as a single number. See estimator. Compare confidence interval; interval estimate.

**Poisson distribution.** A limiting case of the binomial distribution, when the number of trials is large and the common probability is small. The parameter of the approximating Poisson distribution is the number of trials times the common probability, which is the expected number of events. When this number is large, the Poisson distribution may be approximated by a normal distribution.

**population.** Also, universe. All the units of interest to the researcher. Compare sample; sampling frame.

**population size.** Also, size of population. Number of units in the population.

**posterior probability.** See Bayes' rule.

**power.** The probability that a statistical test will reject the null hypothesis. To compute power, one has to fix the size of the test and specify parameter values outside the range given by the null hypothesis. A powerful test has a good chance of detecting an effect when there is an effect to be detected. See beta; significance test. Compare alpha; size;  $p$ -value.

**practical significance.** Substantive importance. Statistical significance does not necessarily establish practical significance. With large samples, small differences can be statistically significant. See significance test.

**practice effects.** Changes in test scores that result from taking the same test twice in succession, or taking two similar tests one after the other.

**predicted value.** See residual.

**predictive validity.** A skills test has predictive validity to the extent that test scores are well correlated with later performance, or more generally with outcomes that the test is intended to predict. See validity. Compare reliability.

**predictor.** See independent variable.

**prior probability.** See Bayes' rule.

**probability.** Chance, on a scale from 0 to 1. Impossibility is represented by 0, certainty by 1. Equivalently, chances may be quoted in percent; 100% corresponds to 1, 5% corresponds to .05, and so forth.

**probability density.** Describes the probability distribution of a random variable. The chance that the random variable falls in an interval equals the area below the density and above the interval. (However, not all random variables have densities.) See probability distribution; random variable.

**probability distribution.** Gives probabilities for possible values or ranges of values of a random variable. Often, the distribution is described in terms of a density. See probability density.

**probability histogram.** See histogram.

**probability model.** Relates probabilities of outcomes to parameters; also, statistical model. The latter connotes unknown parameters.

**probability sample.** A sample drawn from a sampling frame by some objective chance mechanism; each unit has a known probability of being sampled. Such samples minimize selection bias, but can be expensive to draw.

**psychometrics.** The study of psychological measurement and testing.

**qualitative variable; quantitative variable.** Describes qualitative features of subjects in a study (e.g., marital status—never-married, married, widowed, divorced, separated). A quantitative variable describes numerical features of the subjects (e.g., height, weight, income). This is not a hard-and-fast distinction, because qualitative features may be given numerical codes, as with a dummy variable. Quantitative variables may be classified as discrete or continuous. Concepts such as the mean and the standard deviation apply only to quantitative variables. Compare continuous variable; discrete variable; dummy variable. See variable.

**quartile.** The 25th or 75th percentile. See percentile. Compare median.

***R*-squared ( $R^2$ ).** Measures how well a regression equation fits the data. *R*-squared varies between 0 (no fit) and 1 (perfect fit). *R*-squared does not measure the extent to which underlying assumptions are justified. See regression model. Compare multiple correlation coefficient; standard error of regression.

**random error.** Sources of error that are random in their effect, like draws made at random from a box. These are reflected in the error term of a statistical model. Some authors refer to random error as chance error or sampling error. See regression model.

**random variable.** A variable whose possible values occur according to some probability mechanism. For example, if a pair of dice are thrown, the total number of spots is a random variable. The chance of two spots is 1/36, the



chance of three spots is  $2/36$ , and so forth; the most likely number is 7, with chance  $6/36$ .

The expected value of a random variable is the weighted average of the possible values; the weights are the probabilities. In our example, the expected value is

$$\begin{aligned} & \frac{1}{36} \times 2 + \frac{2}{36} \times 3 + \frac{3}{36} \times 4 + \frac{5}{36} \times 6 + \frac{6}{36} \times 7 \\ & + \frac{5}{36} \times 8 + \frac{4}{36} \times 9 + \frac{3}{36} \times 10 + \frac{2}{36} \times 11 + \frac{1}{36} \times 12 \end{aligned}$$

In many problems, the weighted average is computed with respect to the density; then sums must be replaced by integrals. The expected value need not be a possible value for the random variable.

Generally, a random variable will be somewhere around its expected value, but will be off (in either direction) by something like a standard error (SE) or so. If the random variable has a more or less normal distribution, there is about a 68% chance for it to fall in the range expected value – SE to expected value + SE. See normal curve; standard error.

**randomization.** See controlled experiment; randomized controlled experiment.

**randomized controlled experiment.** A controlled experiment in which subjects are placed into the treatment and control groups at random—as if by a lottery. See controlled experiment. Compare observational study.

**range.** The difference between the biggest and the smallest values in a batch of numbers.

**rate.** In an epidemiological study, the number of events, divided by the size of the population; often cross-classified by age and gender. For example, the death rate from heart disease among American men ages 55–64 in 2004 was about three per thousand. Among men ages 65–74, the rate was about seven per thousand. Among women, the rate was about half that for men. Rates adjust for differences in sizes of populations or subpopulations. Often, rates are computed per unit of time, e.g., per thousand persons per year. Data source: Statistical Abstract of the United States tbl. 115 (2008).

**regression coefficient.** The coefficient of a variable in a regression equation. See regression model.

**regression diagnostics.** Procedures intended to check whether the assumptions of a regression model are appropriate.

**regression equation.** See regression model.

**regression line.** The graph of a (simple) regression equation.

**regression model.** A regression model attempts to combine the values of certain variables (the independent or explanatory variables) in order to get expected values for another variable (the dependent variable). Sometimes, the phrase

“regression model” refers to a probability model for the data; if no qualifications are made, the model will generally be linear, and errors will be assumed independent across observations, with common variance. The coefficients in the linear combination are called regression coefficients; these are parameters. At times, “regression model” refers to an equation (“the regression equation”) estimated from data, typically by least squares.

For example, in a regression study of salary differences between men and women in a firm, the analyst may include a dummy variable for gender, as well as statistical controls such as education and experience to adjust for productivity differences between men and women. The dummy variable would be defined as 1 for the men and 0 for the women. Salary would be the dependent variable; education, experience, and the dummy would be the independent variables. See least squares; multiple regression; random error; variance. Compare general linear model.

**relative frequency.** See frequency.

**relative risk.** A measure of association used in epidemiology. For example, if 10% of all people exposed to a chemical develop a disease, compared to 5% of people who are not exposed, then the disease occurs twice as frequently among the exposed people: The relative risk is  $10\%/5\% = 2$ . A relative risk of 1 indicates no association. For more details, see Leon Gordis, *Epidemiology* (4th ed. 2008). Compare odds ratio.

**reliability.** The extent to which a measurement process gives the same results on repeated measurement of the same thing. Compare validity.

**representative sample.** Not a well-defined technical term. A sample judged to fairly represent the population, or a sample drawn by a process likely to give samples that fairly represent the population, for example, a large probability sample.

**resampling.** See bootstrap.

**residual.** The difference between an actual and a predicted value. The predicted value comes typically from a regression equation, and is better called the fitted value, because there is no real prediction going on. See regression model; independent variable.

**response variable.** See independent variable.

**risk.** Expected loss. “Expected” means on average, over the various datasets that could be generated by the statistical model under examination. Usually, risk cannot be computed exactly but has to be estimated, because the parameters in the statistical model are unknown and must be estimated. See loss function; random variable.

**risk factor.** See independent variable.

**robust.** A statistic or procedure that does not change much when data or assumptions are modified slightly.

**sample.** A set of units collected for study. Compare population.

**sample size.** Also, size of sample. The number of units in a sample.

**sample weights.** See stratified random sample.

**sampling distribution.** The distribution of the values of a statistic, over all possible samples from a population. For example, suppose a random sample is drawn. Some values of the sample mean are more likely; others are less likely. The sampling distribution specifies the chance that the sample mean will fall in one interval rather than another.

**sampling error.** A sample is part of a population. When a sample is used to estimate a numerical characteristic of the population, the estimate is likely to differ from the population value because the sample is not a perfect microcosm of the whole. If the estimate is unbiased, the difference between the estimate and the exact value is sampling error. More generally,

$$\text{estimate} = \text{true value} + \text{bias} + \text{sampling error}$$

Sampling error is also called chance error or random error. See standard error. Compare bias; nonsampling error.

**sampling frame.** A list of units designed to represent the entire population as completely as possible. The sample is drawn from the frame.

**sampling interval.** See systematic sample.

**scatter diagram.** Also, scatterplot; scattergram. A graph showing the relationship between two variables in a study. Each dot represents one subject. One variable is plotted along the horizontal axis, the other variable is plotted along the vertical axis. A scatter diagram is homoscedastic when the spread is more or less the same inside any vertical strip. If the spread changes from one strip to another, the diagram is heteroscedastic.

**selection bias.** Systematic error due to nonrandom selection of subjects for study.

**sensitivity.** In clinical medicine, the probability that a test for a disease will give a positive result given that the patient has the disease. Sensitivity is analogous to the power of a statistical test. Compare specificity.

**sensitivity analysis.** Analyzing data in different ways to see how results depend on methods or assumptions.

**sign test.** A statistical test based on counting and the binomial distribution. For example, a Finnish study of twins found 22 monozygotic twin pairs where 1 twin smoked, 1 did not, and at least 1 of the twins had died. That sets up a race to death. In 17 cases, the smoker died first; in 5 cases, the nonsmoker died first. The null hypothesis is that smoking does not affect time to death, so the chances are 50-50 for the smoker to die first. On the null hypothesis, the chance that the smoker will win the race 17 or more times out of 22 is

8/1000. That is the  $p$ -value. The  $p$ -value can be computed from the binomial distribution. For additional detail, see Michael O. Finkelstein & Bruce Levin, *Statistics for Lawyers* 339–41 (2d ed. 2001); David A. Freedman et al., *Statistics* 262–63 (4th ed. 2007).

**significance level.** See fixed significance level;  $p$ -value.

**significance test.** Also, statistical test; hypothesis test; test of significance. A significance test involves formulating a statistical hypothesis and a test statistic, computing a  $p$ -value, and comparing  $p$  to some preestablished value ( $\alpha$ ) to decide if the test statistic is significant. The idea is to see whether the data conform to the predictions of the null hypothesis. Generally, a large test statistic goes with a small  $p$ -value; and small  $p$ -values would undermine the null hypothesis.

For example, suppose that a random sample of male and female employees were given a skills test and the mean scores of the men and women were different—in the sample. To judge whether the difference is due to sampling error, a statistician might consider the implications of competing hypotheses about the difference in the population. The null hypothesis would say that on average, in the population, men and women have the same scores: The difference observed in the data is then just due to sampling error. A one-sided alternative hypothesis would be that on average, in the population, men score higher than women. The one-sided test would reject the null hypothesis if the sample men score substantially higher than the women—so much so that the difference is hard to explain on the basis of sampling error.

In contrast, the null hypothesis could be tested against the two-sided alternative that on average, in the population, men score differently than women—higher or lower. The corresponding two-sided test would reject the null hypothesis if the sample men score substantially higher or substantially lower than the women.

The one-sided and two-sided tests would both be based on the same data, and use the same  $t$ -statistic. However, if the men in the sample score higher than the women, the one-sided test would give a  $p$ -value only half as large as the two-sided test; that is, the one-sided test would appear to give stronger evidence against the null hypothesis. (“One-sided” and “one-tailed” are synonymous; so are “two-sided and “two-tailed.”) See  $p$ -value; statistical hypothesis;  $t$ -statistic.

**significant.** See  $p$ -value; practical significance; significance test.

**simple random sample.** A random sample in which each unit in the sampling frame has the same chance of being sampled. The investigators take a unit at random (as if by lottery), set it aside, take another at random from what is left, and so forth.

**simple regression.** A regression equation that includes only one independent variable. Compare multiple regression.

**size.** A synonym for alpha ( $\alpha$ ).

**skip factor.** See systematic sample.

**specificity.** In clinical medicine, the probability that a test for a disease will give a negative result given that the patient does not have the disease. Specificity is analogous to  $1 - \alpha$ , where  $\alpha$  is the significance level of a statistical test. Compare sensitivity.

**spurious correlation.** When two variables are correlated, one is not necessarily the cause of the other. The vocabulary and shoe size of children in elementary school, for example, are correlated—but learning more words will not make the feet grow. Such noncausal correlations are said to be spurious. (Originally, the term seems to have been applied to the correlation between two rates with the same denominator: Even if the numerators are unrelated, the common denominator will create some association.) Compare confounding variable.

**standard deviation (SD).** Indicates how far a typical element deviates from the average. For example, in round numbers, the average height of women age 18 and over in the United States is 5 feet 4 inches. However, few women are exactly average; most will deviate from average, at least by a little. The SD is sort of an average deviation from average. For the height distribution, the SD is 3 inches. The height of a typical woman is around 5 feet 4 inches, but is off that average value by something like 3 inches.

For distributions that follow the normal curve, about 68% of the elements are in the range from 1 SD below the average to 1 SD above the average. Thus, about 68% of women have heights in the range 5 feet 1 inch to 5 feet 7 inches. Deviations from the average that exceed 3 or 4 SDs are extremely unusual. Many authors use standard deviation to also mean standard error. See standard error.

**standard error (SE).** Indicates the likely size of the sampling error in an estimate. Many authors use the term standard deviation instead of standard error. Compare expected value; standard deviation.

**standard error of regression.** Indicates how actual values differ (in some average sense) from the fitted values in a regression model. See regression model; residual. Compare *R*-squared.

**standard normal.** See normal distribution.

**standardization.** See standardized variable.

**standardized variable.** Transformed to have mean zero and variance one. This involves two steps: (1) subtract the mean; (2) divide by the standard deviation.

**statistic.** A number that summarizes data. A statistic refers to a sample; a parameter or a true value refers to a population or a probability model.

**statistical controls.** Procedures that try to filter out the effects of confounding variables on non-experimental data, for example, by adjusting through statistical procedures such as multiple regression. Variables in a multiple regression

equation. See multiple regression; confounding variable; observational study. Compare controlled experiment.

**statistical dependence.** See dependence.

**statistical hypothesis.** Generally, a statement about parameters in a probability model for the data. The null hypothesis may assert that certain parameters have specified values or fall in specified ranges; the alternative hypothesis would specify other values or ranges. The null hypothesis is tested against the data with a test statistic; the null hypothesis may be rejected if there is a statistically significant difference between the data and the predictions of the null hypothesis.

Typically, the investigator seeks to demonstrate the alternative hypothesis; the null hypothesis would explain the findings as a result of mere chance, and the investigator uses a significance test to rule out that possibility. See significance test.

**statistical independence.** See independence.

**statistical model.** See probability model.

**statistical test.** See significance test.

**statistical significance.** See  $p$ -value.

**stratified random sample.** A type of probability sample. The researcher divides the population into relatively homogeneous groups called “strata,” and draws a random sample separately from each stratum. Dividing the population into strata is called “stratification.” Often the sampling fraction will vary from stratum to stratum. Then sampling weights should be used to extrapolate from the sample to the population. For example, if 1 unit in 10 is sampled from stratum A while 1 unit in 100 is sampled from stratum B, then each unit drawn from A counts as 10, and each unit drawn from B counts as 100. The first kind of unit has weight 10; the second has weight 100. See Freedman et al., *Statistics* 401 (4th ed. 2007).

**stratification.** See independent variable; stratified random sample.

**study validity.** See validity.

**subjectivist.** See Bayesian.

**systematic error.** See bias.

**systematic sample.** Also, list sample. The elements of the population are numbered consecutively as 1, 2, 3, . . . . The investigators choose a starting point and a “sampling interval” or “skip factor”  $k$ . Then, every  $k$ th element is selected into the sample. If the starting point is 1 and  $k = 10$ , for example, the sample would consist of items 1, 11, 21, . . . . Sometimes the starting point is chosen at random from 1 to  $k$ : this is a random-start systematic sample.

**$t$ -statistic.** A test statistic, used to make the  $t$ -test. The  $t$ -statistic indicates how far away an estimate is from its expected value, relative to the standard error. The expected value is computed using the null hypothesis that is being tested.

Some authors refer to the *t*-statistic, others to the *z*-statistic, especially when the sample is large. With a large sample, a *t*-statistic larger than 2 or 3 in absolute value makes the null hypothesis rather implausible—the estimate is too many standard errors away from its expected value. See statistical hypothesis; significance test; *t*-test.

***t*-test.** A statistical test based on the *t*-statistic. Large *t*-statistics are beyond the usual range of sampling error. For example, if *t* is bigger than 2, or smaller than  $-2$ , then the estimate is statistically significant at the 5% level; such values of *t* are hard to explain on the basis of sampling error. The scale for *t*-statistics is tied to areas under the normal curve. For example, a *t*-statistic of 1.5 is not very striking, because  $13\% = 13/100$  of the area under the normal curve is outside the range from  $-1.5$  to  $1.5$ . On the other hand,  $t = 3$  is remarkable: Only  $3/1000$  of the area lies outside the range from  $-3$  to  $3$ . This discussion is predicated on having a reasonably large sample; in that context, many authors refer to the *z*-test rather than the *t*-test.

Consider testing the null hypothesis that the average of a population equals a given value; the population is known to be normal. For small samples, the *t*-statistic follows Student's *t*-distribution (when the null hypothesis holds) rather than the normal curve; larger values of *t* are required to achieve significance. The relevant *t*-distribution depends on the number of degrees of freedom, which in this context equals the sample size minus one. A *t*-test is not appropriate for small samples drawn from a population that is not normal. See *p*-value; significance test; statistical hypothesis.

**test statistic.** A statistic used to judge whether data conform to the null hypothesis. The parameters of a probability model determine expected values for the data; differences between expected values and observed values are measured by a test statistic. Such test statistics include the chi-squared statistic ( $\chi^2$ ) and the *t*-statistic. Generally, small values of the test statistic are consistent with the null hypothesis; large values lead to rejection. See *p*-value; statistical hypothesis; *t*-statistic.

**time series.** A series of data collected over time, for example, the Gross National Product of the United States from 1945 to 2005.

**treatment group.** See controlled experiment.

**two-sided hypothesis; two-tailed hypothesis.** An alternative hypothesis asserting that the values of a parameter are different from—either greater than or less than—the value asserted in the null hypothesis. A two-sided alternative hypothesis suggests a two-sided (or two-tailed) test. See significance test; statistical hypothesis. Compare one-sided hypothesis.

**two-sided test; two-tailed test.** See two-sided hypothesis.

**Type I error.** A statistical test makes a Type I error when (1) the null hypothesis is true and (2) the test rejects the null hypothesis, i.e., there is a false posi-

tive. For example, a study of two groups may show some difference between samples from each group, even when there is no difference in the population. When a statistical test deems the difference to be significant in this situation, it makes a Type I error. See significance test; statistical hypothesis. Compare alpha; Type II error.

**Type II error.** A statistical test makes a Type II error when (1) the null hypothesis is false and (2) the test fails to reject the null hypothesis, i.e., there is a false negative. For example, there may not be a significant difference between samples from two groups when, in fact, the groups are different. See significance test; statistical hypothesis. Compare beta; Type I error.

**unbiased estimator.** An estimator that is correct on average, over the possible datasets. The estimates have no systematic tendency to be high or low. Compare bias.

**uniform distribution.** For example, a whole number picked at random from 1 to 100 has the uniform distribution: All values are equally likely. Similarly, a uniform distribution is obtained by picking a real number at random between 0.75 and 3.25: The chance of landing in an interval is proportional to the length of the interval.

**validity.** Measurement validity is the extent to which an instrument measures what it is supposed to, rather than something else. The validity of a standardized test is often indicated by the correlation coefficient between the test scores and some outcome measure (the criterion variable). See content validity; differential validity; predictive validity. Compare reliability.

Study validity is the extent to which results from a study can be relied upon. Study validity has two aspects, internal and external. A study has high internal validity when its conclusions hold under the particular circumstances of the study. A study has high external validity when its results are generalizable. For example, a well-executed randomized controlled double-blind experiment performed on an unusual study population will have high internal validity because the design is good; but its external validity will be debatable because the study population is unusual.

Validity is used also in its ordinary sense: assumptions are valid when they hold true for the situation at hand.

**variable.** A property of units in a study, which varies from one unit to another, for example, in a study of households, household income; in a study of people, employment status (employed, unemployed, not in labor force).

**variance.** The square of the standard deviation. Compare standard error; covariance.

**weights.** See stratified random sample.

**within-observer variability.** Differences that occur when an observer measures the same thing twice, or measures two things that are virtually the same. Compare between-observer variability.



**z-statistic.** See *t*-statistic.

**z-test.** See *t*-test.

## References on Statistics

### *General Surveys*

David Freedman et al., *Statistics* (4th ed. 2007).

Darrell Huff, *How to Lie with Statistics* (1993).

Gregory A. Kimble, *How to Use (and Misuse) Statistics* (1978).

David S. Moore & William I. Notz, *Statistics: Concepts and Controversies* (2005).

Michael Oakes, *Statistical Inference: A Commentary for the Social and Behavioral Sciences* (1986).

*Statistics: A Guide to the Unknown* (Roxy Peck et al. eds., 4th ed. 2005).

Hans Zeisel, *Say It with Figures* (6th ed. 1985).

### *Reference Works for Lawyers and Judges*

David C. Baldus & James W.L. Cole, *Statistical Proof of Discrimination* (1980 & Supp. 1987) (continued as Ramona L. Paetzold & Steven L. Willborn, *The Statistics of Discrimination: Using Statistical Evidence in Discrimination Cases* (1994) (updated annually)).

David W. Barnes & John M. Conley, *Statistical Evidence in Litigation: Methodology, Procedure, and Practice* (1986 & Supp. 1989).

James Brooks, *A Lawyer's Guide to Probability and Statistics* (1990).

Michael O. Finkelstein & Bruce Levin, *Statistics for Lawyers* (2d ed. 2001).

*Modern Scientific Evidence: The Law and Science of Expert Testimony* (David L. Faigman et al. eds., Volumes 1 and 2, 2d ed. 2002) (updated annually).

David H. Kaye et al., *The New Wigmore: A Treatise on Evidence: Expert Evidence* § 12 (2d ed. 2011) (updated annually).

National Research Council, *The Evolving Role of Statistical Assessments as Evidence in the Courts* (Stephen E. Fienberg ed., 1989).

*Statistical Methods in Discrimination Litigation* (David H. Kaye & Mikel Aickin eds., 1986).

Hans Zeisel & David Kaye, *Prove It with Figures: Empirical Methods in Law and Litigation* (1997).

### *General Reference*

*Encyclopedia of Statistical Sciences* (Samuel Kotz et al. eds., 2d ed. 2005).

# Exhibit 54

# On the Impossibility of Informationally Efficient Markets

By SANFORD J. GROSSMAN AND JOSEPH E. STIGLITZ\*

If competitive equilibrium is defined as a situation in which prices are such that all arbitrage profits are eliminated, is it possible that a competitive economy always be in equilibrium? Clearly not, for then those who arbitrage make no (private) return from their (privately) costly activity. Hence the assumptions that all markets, including that for information, are always in equilibrium and always perfectly arbitrated are inconsistent when arbitrage is costly.

We propose here a model in which there is an equilibrium degree of disequilibrium: prices reflect the information of informed individuals (arbitrageurs) but only partially, so that those who expend resources to obtain information do receive compensation. How informative the price system is depends on the number of individuals who are informed; but the number of individuals who are informed is itself an endogenous variable in the model.

The model is the simplest one in which prices perform a well-articulated role in conveying information from the informed to the uninformed. When informed individuals observe information that the return to a security is going to be high, they bid its price up, and conversely when they observe information that the return is going to be low. Thus the price system makes publicly available the information obtained by informed individuals to the uninformed. In general, however, it does this imperfectly; this is perhaps lucky, for were it to do it perfectly, an equilibrium would not exist.

In the introduction, we shall discuss the general methodology and present some con-

jectures concerning certain properties of the equilibrium. The remaining analytic sections of the paper are devoted to analyzing in detail an important example of our general model, in which our conjectures concerning the nature of the equilibrium can be shown to be correct. We conclude with a discussion of the implications of our approach and results, with particular emphasis on the relationship of our results to the literature on "efficient capital markets."

## I. The Model

Our model can be viewed as an extension of the noisy rational expectations model introduced by Robert Lucas and applied to the study of information flows between traders by Jerry Green (1973); Grossman (1975, 1976, 1978); and Richard Kihlstrom and Leonard Mirman. There are two assets: a safe asset yielding a return  $R$ , and a risky asset, the return to which,  $u$ , varies randomly from period to period. The variable  $u$  consists of two parts,

$$(1) \quad u = \theta + \varepsilon$$

where  $\theta$  is observable at a cost  $c$ , and  $\varepsilon$  is unobservable.<sup>1</sup> Both  $\theta$  and  $\varepsilon$  are random variables. There are two types of individuals, those who observe  $\theta$  (informed traders), and those who observe only price (uninformed traders). In our simple model, all individuals are, *ex ante*, identical; whether they are informed or uninformed just depends on whether they have spent  $c$  to obtain information. Informed traders' demands will depend on  $\theta$  and the price of the risky asset  $P$ . Uninformed traders' demands

\*University of Pennsylvania and Princeton University, respectively. Research support under National Science Foundation grants SOC76-18771 and SOC77-15980 is gratefully acknowledged. This is a revised version of a paper presented at the Econometric Society meetings, Winter 1975, at Dallas, Texas.

<sup>1</sup>An alternative interpretation is that  $\theta$  is a "measurement" of  $u$  with error. The mathematics of this alternative interpretation differ slightly, but the results are identical.

will depend only on  $P$ , but we shall assume that they have rational expectations; they learn the relationship between the distribution of return and the price, and use this in deriving their demand for the risky assets. If  $x$  denotes the supply of the risky asset, an equilibrium when a given percentage,  $\lambda$ , of traders are informed, is thus a price function  $P_\lambda(\theta, x)$  such that, when demands are formulated in the way described, demand equals supply. We assume that uninformed traders do not observe  $x$ . Uninformed traders are prevented from learning  $\theta$  via observations of  $P_\lambda(\theta, x)$  because they cannot distinguish variations in price due to changes in the informed trader's information from variations in price due to changes in aggregate supply. Clearly,  $P_\lambda(\theta, x)$  reveals some of the informed trader's information to the uninformed traders.

We can calculate the expected utility of the informed and the expected utility of the uninformed. If the former is greater than the latter (taking account of the cost of information), some individuals switch from being uninformed to being informed (and conversely). An overall equilibrium requires the two to have the same expected utility. As more individuals become informed, the expected utility of the informed falls relative to the uninformed for two reasons:

(a) The price system becomes more informative because variations in  $\theta$  have a greater effect on aggregate demand and thus on price when more traders observe  $\theta$ . Thus, more of the information of the informed is available to the uninformed. Moreover, the informed gain more from trade with the uninformed than do the uninformed. The informed, on average, buy securities when they are "underpriced" and sell them when they are "overpriced" (relative to what they would have been if information were equalized).<sup>2</sup> As the price system becomes more informative, the difference in their information—and hence the magnitude by

which the informed can gain relative to the uninformed—is reduced.

(b) Even if the above effect did not occur, the increase in the ratio of informed to uninformed means that the relative gains of the informed, on a per capita basis, in trading with the uninformed will be smaller.

We summarize the above characterization of the equilibrium of the economy in the following two conjectures:

*Conjecture 1:* The more individuals who are informed, the more informative is the price system.

*Conjecture 2:* The more individuals who are informed, the lower the ratio of expected utility of the informed to the uninformed.

(Conjecture 1 obviously requires a definition of "more informative"; this is given in the next section and in fn. 7.)

The equilibrium number of informed and uninformed individuals in the economy will depend on a number of critical parameters: the cost of information, how informative the price system is (how much noise there is to interfere with the information conveyed by the price system), and how informative the information obtained by an informed individual is.

*Conjecture 3:* The higher the cost of information, the smaller will be the equilibrium percentage of individuals who are informed.

*Conjecture 4:* If the quality of the informed trader's information increases, the more their demands will vary with their information and thus the more prices will vary with  $\theta$ . Hence, the price system becomes more informative. The equilibrium proportion of informed to uninformed may be either increased or decreased, because even though the value of being informed has increased due to the increased quality of  $\theta$ , the value of being uninformed has also increased because the price system becomes more informative.

*Conjecture 5:* The greater the magnitude of noise, the less informative will the price system be, and hence the lower the expected utility of uninformed individuals. Hence, in equilibrium the greater the magnitude of noise, the larger the proportion of informed individuals.

<sup>2</sup>The framework described herein does not explicitly model the effect of variations in supply, i.e.,  $x$  on commodity storage. The effect of futures markets and storage capabilities on the informativeness of the price system was studied by Grossman (1975, 1977).

*Conjecture 6:* In the limit, when there is no noise, prices convey all information, and there is no incentive to purchase information. Hence, the only possible equilibrium is one with no information. But if everyone is uninformed, it clearly pays some individual to become informed.<sup>3</sup> Thus, there does not exist a competitive equilibrium.<sup>4</sup>

Trade among individuals occurs either because tastes (risk aversions) differ, endowments differ, or beliefs differ. This paper focuses on the last of these three. An interesting feature of the equilibrium is that beliefs may be precisely identical in either one of two situations: when all individuals are informed or when all individuals are uninformed. This gives rise to:

*Conjecture 7:* That, other things being equal, markets will be thinner under those conditions in which the percentage of individuals who are informed ( $\lambda$ ) is either near zero or near unity. For example, markets will be thin when there is very little noise in the system (so  $\lambda$  is near zero), or when costs of information are very low (so  $\lambda$  is near unity).

In the last few paragraphs, we have provided a number of conjectures describing the nature of the equilibrium when prices convey information. Unfortunately, we have not been able to obtain a general proof of any of these propositions. What we have been able to do is to analyze in detail an interesting example, entailing constant absolute risk-aversion utility functions and normally distributed random variables. In this example, the equilibrium price distribution can actually be calculated, and all of

the conjectures provided above can be verified. The next sections are devoted to solving for the equilibrium in this particular example.<sup>5</sup>

## II. Constant Absolute Risk-Aversion Model

### A. The Securities

The  $i$ th trader is assumed to be endowed with stocks of two types of securities:  $\bar{M}_i$ , the riskless asset, and  $\bar{X}_i$ , a risky asset. Let  $P$  be the current price of risky assets and set the price of risk free assets equal to unity. The  $i$ th trader's budget constraint is

$$(2) \quad PX_i + M_i = W_{0i} \equiv \bar{M}_i + P\bar{X}_i$$

Each unit of the risk free asset pays  $R$  "dollars" at the end of the period, while each unit of the risky asset pays  $u$  dollars. If at the end of the period, the  $i$ th trader holds a portfolio  $(M_i, X_i)$ , his wealth will be

$$(3) \quad W_{1i} = RM_i + uX_i$$

### B. Individual's Utility Maximization

Each individual has a utility function  $V_i(W_{1i})$ . For simplicity, we assume all individuals have the same utility function and so drop the subscripts  $i$ . Moreover, we assume the utility function is exponential, i.e.,

$$V(W_{1i}) = -e^{-aW_{1i}}, \quad a > 0$$

where  $a$  is the coefficient of absolute risk aversion. Each trader desires to maximize expected utility, using whatever information is available to him, and to decide on what information to acquire on the basis of the consequences to his expected utility.

Assume that in equation (1)  $\theta$  and  $\epsilon$  have a multivariate normal distribution, with

$$(4) \quad E\epsilon = 0$$

$$(5) \quad E\theta\epsilon = 0$$

$$(6) \quad \text{Var}(u^*|\theta) = \text{Var}\epsilon^* \equiv \sigma_\epsilon^2 > 0$$

<sup>3</sup>That is, with no one informed, an individual can only get information by paying  $c$  dollars, since no information is revealed by the price system. By paying  $c$  dollars an individual will be able to predict better than the market when it is optimal to hold the risky asset as opposed to the risk-free asset. Thus his expected utility will be higher than an uninformed person gross of information costs. Thus for  $c$  sufficiently low all uninformed people will desire to be informed.

<sup>4</sup>See Grossman (1975, 1977) for a formal example of this phenomenon in futures markets. See Stiglitz (1971, 1974) for a general discussion of information and the possibility of nonexistence of equilibrium in capital markets.

<sup>5</sup>The informational equilibria discussed here may not, in general, exist. See Green (1977). Of course, for the utility function we choose equilibrium does exist.

since  $\theta$  and  $\varepsilon$  are uncorrelated. Throughout this paper we will put a \* above a symbol to emphasize that it is a random variable. Since  $W_{it}$  is a linear function of  $\varepsilon$ , for a given portfolio allocation, and a linear function of a normally distributed random variable is normally distributed, it follows that  $W_{it}$  is normal conditional on  $\theta$ . Then, using (2) and (3) the expected utility of the *informed* trader with information  $\theta$  can be written

$$\begin{aligned} (7) \quad E(V(W_{it}^*)|\theta) &= \\ &= -\exp\left(-a\left\{E[W_{it}^*|\theta] - \frac{a}{2}\text{Var}[W_{it}^*|\theta]\right\}\right) \\ &= -\exp\left(-a\left[RW_{0t} + X_I\{E(u^*|\theta) - RP\} \right. \right. \\ &\quad \left. \left. - \frac{a}{2}X_I^2\text{Var}(u^*|\theta)\right]\right) \\ &= -\exp\left(-a\left[RW_{0t} + X_I(\theta - RP) \right. \right. \\ &\quad \left. \left. - \frac{a}{2}X_I^2\sigma_\varepsilon^2\right]\right) \end{aligned}$$

where  $X_I$  is an informed individual's demand for the risky security. Maximizing (7) with respect to  $X_I$  yields a demand function for risky assets:

$$(8) \quad X_I(P, \theta) = \frac{\theta - RP}{a\sigma_\varepsilon^2}$$

The right-hand side of (8) shows the familiar result that with constant absolute risk aversion, a trader's demand does not depend on wealth; hence the subscript  $i$  is not on the left-hand side of (8).

We now derive the demand function for the uninformed. Let us assume the only source of "noise" is the per capita supply of the risky security  $x$ .

Let  $P^*(\cdot)$  be some particular price function of  $(\theta, x)$  such that  $u^*$  and  $P^*$  are jointly normally distributed. (We will prove that this exists below.)

Then, we can write for the uninformed individual

$$\begin{aligned} (7') \quad E(V(W_{it}^*)|P^*) &= -\exp\left[-a\left\{E[W_{it}^*|P^*] \right. \right. \\ &\quad \left. \left. - \frac{a}{2}\text{Var}[W_{it}^*|P^*]\right\}\right] \\ &= -\exp\left[-a\left\{RW_{0t} + X_U(E[u^*|P^*] - RP) \right. \right. \\ &\quad \left. \left. - \frac{a}{2}X_U^2\text{Var}[u^*|P^*]\right\}\right] \end{aligned}$$

The demands of the uninformed will thus be a function of the price function  $P^*$  and the actual price  $P$ .

$$\begin{aligned} (8') \quad X_U(P; P^*) &= \\ &= \frac{E[u^*|P^*(\theta, x) = P] - RP}{a\text{Var}[u^*|P^*(\theta, x) = P]} \end{aligned}$$

### C. Equilibrium Price Distribution

If  $\lambda$  is some particular fraction of traders who decide to become informed, then define an equilibrium price system as a function of  $(\theta, x)$ ,  $P_\lambda(\theta, x)$ , such that for all  $(\theta, x)$  per capita demands for the risky assets equal supplies:

$$\begin{aligned} (9) \quad \lambda X_I(P_\lambda(\theta, x), \theta) &+ \\ &+ (1 - \lambda)X_U(P_\lambda(\theta, x); P_\lambda^*) = x \end{aligned}$$

The function  $P_\lambda(\theta, x)$  is a statistical equilibrium in the following sense. If over time uninformed traders observe many realizations of  $(u^*, P_\lambda^*)$ , then they learn the joint distribution of  $(u^*, P_\lambda^*)$ . After all learning about the joint distribution of  $(u^*, P_\lambda^*)$  ceases, all traders will make allocations and form expectations such that this joint distribution persists over time. This follows from (8), (8'), and (9), where the market-clearing price that comes about is the one which takes into account the fact that uninformed traders have learned that it contains information.



We shall now prove that there exists an equilibrium price distribution such that  $P^*$  and  $u^*$  are jointly normal. Moreover, we shall be able to characterize the price distribution. We define

$$(10a) \quad w_\lambda(\theta, x) = \theta - \frac{a\sigma_\epsilon^2}{\lambda}(x - Ex^*)$$

for  $\lambda > 0$ , and define  $w_0(\theta, x)$  as the number:

$$(10b) \quad w_0(\theta, x) = x \quad \text{for all } (\theta, x)$$

where  $w_\lambda$  is just the random variable  $\theta$ , plus noise.<sup>6</sup> The magnitude of the noise is inversely proportional to the proportion of informed traders, but is proportional to the variance of  $\epsilon$ . We shall prove that the equilibrium price is just a linear function of  $w_\lambda$ . Thus, if  $\lambda > 0$ , the price system conveys information about  $\theta$ , but it does so imperfectly.

#### D. Existence of Equilibrium and a Characterization Theorem

**THEOREM 1:** *If  $(\theta^*, \epsilon^*, x^*)$  has a nondegenerate joint normal distribution such that  $\theta^*$ ,  $\epsilon^*$ , and  $x^*$  are mutually independent, then there exists a solution to (9) which has the form  $P_\lambda(\theta, x) = \alpha_1 + \alpha_2 w_\lambda(\theta, x)$ , where  $\alpha_1$  and  $\alpha_2$  are real numbers which may depend on  $\lambda$ , such that  $\alpha_2 > 0$ . (If  $\lambda = 0$ , the price contains no information about  $\theta$ .) The exact form of  $P_\lambda(\theta, x)$  is given in equation (A10) in Appendix B. The proof of this theorem is also in Appendix B.*

The importance of Theorem 1 rests in the simple characterization of the information in the equilibrium price system:  $P_\lambda^*$  is informationally equivalent to  $w_\lambda^*$ . From (10)  $w_\lambda^*$  is a "mean-preserving spread" of  $\theta$ ; i.e.,  $E[w_\lambda^*|\theta] = \theta$  and

$$(11) \quad \text{Var}[w_\lambda^*|\theta] = \frac{a^2\sigma_\epsilon^4}{\lambda^2} \text{Var} x^*$$

<sup>6</sup>If  $y' = y + Z$ , and  $E[Z|y] = 0$ , then  $y'$  is just  $y$  plus noise.

For each replication of the economy,  $\theta$  is the information that uninformed traders would like to know. But the noise  $x^*$  prevents  $w_\lambda^*$  from revealing  $\theta$ . How well-informed uninformed traders can become from observing  $P_\lambda^*$  (equivalently  $w_\lambda^*$ ) is measured by  $\text{Var}[w_\lambda^*|\theta]$ . When  $\text{Var}[w_\lambda^*|\theta]$  is zero,  $w_\lambda^*$  and  $\theta$  are perfectly correlated. Hence when uninformed firms observe  $w_\lambda^*$ , this is equivalent to observing  $\theta$ . On the other hand, when  $\text{Var}[w_\lambda^*|\theta]$  is very large, there are "many" realizations of  $w_\lambda^*$  that are associated with a given  $\theta$ . In this case the observation of a particular  $w_\lambda^*$  tells very little about the actual  $\theta$  which generated it.<sup>7</sup>

From equation (11) it is clear that large noise (high  $\text{Var} x^*$ ) leads to an imprecise price system. The other factor which determines the precision of the price system ( $a^2\sigma_\epsilon^4/\lambda^2$ ) is more subtle. When  $a$  is small (the individual is not very risk averse) or  $\sigma_\epsilon^2$  is small (the information is very precise), an informed trader will have a demand for risky assets which is very responsive to changes in  $\theta$ . Further, the larger  $\lambda$  is, the more responsive is the total demand of informed traders. Thus small ( $a^2\sigma_\epsilon^4/\lambda^2$ ) means that the aggregate demand of informed traders is very responsive to  $\theta$ . For a fixed amount of noise (i.e., fixed  $\text{Var} x^*$ ) the larger are the movements in aggregate demand which are due to movements in  $\theta$ , the more will price movements be due to movements in  $\theta$ . That is,  $x^*$  becomes less important relative to  $\theta$  in determining price movements. Therefore, for small ( $a^2\sigma_\epsilon^4/\lambda^2$ ) uninformed traders are able to confidently know that price is, for example, unusually high due to  $\theta$  being high. In this way information from informed traders is transferred to uninformed traders.

<sup>7</sup>Formally,  $w_\lambda^*$  is an experiment in the sense of Blackwell which gives information about  $\theta$ . It is easy to show that, *ceteris paribus*, the smaller  $\text{Var}(w_\lambda^*|\theta)$  the more "informative" (or sufficient) in the sense of Blackwell, is the experiment; see Grossman, Kihlstrom, and Mirman (p. 539).

E. *Equilibrium in the Information Market*

What we have characterized so far is the equilibrium price distribution for given  $\lambda$ . We now define an *overall* equilibrium to be a pair  $(\lambda, P_\lambda^*)$  such that the expected utility of the informed is equal to that of the uninformed if  $0 < \lambda < 1$ ;  $\lambda = 0$  if the expected utility of the informed is less than that of the uninformed at  $P_0^*$ ;  $\lambda = 1$  if the expected utility of the informed is greater than the uninformed at  $P_1^*$ . Let

$$(12a) \quad W_{ii}^\lambda \equiv R(W_{0i} - c) \\ + [u - RP_\lambda(\theta, x)]X_i(P_\lambda(\theta, x), \theta)$$

$$(12b) \quad W_{0i}^\lambda \equiv RW_{0i} \\ + [u - RP_\lambda(\theta, x)]X_U(P_\lambda(\theta, x); P_\lambda^*)$$

where  $c$  is the cost of observing a realization of  $\theta^*$ . Equation (12a) gives the end of period wealth of a trader if he decides to become informed, while (12b) gives his wealth if he decides to be uninformed. Note that end of period wealth is random due to the randomness of  $W_{0i}$ ,  $u$ ,  $\theta$ , and  $x$ .

In evaluating the expected utility of  $W_{ii}^\lambda$ , we do not assume that a trader knows which realization of  $\theta^*$  he gets to observe if he pays  $c$  dollars. A trader pays  $c$  dollars and then gets to observe some realization of  $\theta^*$ . The overall expected utility of  $W_{ii}^\lambda$  averages over all possible  $\theta^*$ ,  $\varepsilon^*$ ,  $x^*$ , and  $W_{0i}$ . The variable  $W_{0i}$  is random for two reasons. First from (2) it depends on  $P_\lambda(\theta, x)$ , which is random as  $(\theta, x)$  is random. Secondly, in what follows we will assume that  $X_i$  is random.

We will show below that  $EV(W_{ii}^\lambda)/EV(W_{0i}^\lambda)$  is independent of  $i$ , but is a function of  $\lambda$ ,  $a$ ,  $c$ , and  $\sigma_\varepsilon^2$ . More precisely, in Appendix B we prove

**THEOREM 2:** *Under the assumptions of Theorem 1, and if  $\bar{X}_i$  is independent of  $(u^*, \theta^*, x^*)$  then*

$$(13) \quad \frac{EV(W_{ii}^\lambda)}{EV(W_{0i}^\lambda)} = e^{ac} \sqrt{\frac{Var(u^*|\theta)}{Var(u^*|w_\lambda)}}$$

F. *Existence of Overall Equilibrium*

Theorem 2 is useful, both in proving the uniqueness of overall equilibrium and in analyzing comparative statics. Overall equilibrium, it will be recalled, requires that for  $0 < \lambda < 1$ ,  $EV(W_{ii}^\lambda)/EV(W_{0i}^\lambda) = 1$ . But from (13)

$$(14) \quad \frac{EV(W_{ii}^\lambda)}{EV(W_{0i}^\lambda)} \\ = e^{ac} \sqrt{\frac{Var(u^*|\theta)}{Var(u^*|w_\lambda)}} \equiv \gamma(\lambda)$$

Hence overall equilibrium simply requires, for  $0 < \lambda < 1$ ,

$$(15) \quad \gamma(\lambda) = 1$$

More precisely, we now prove

**THEOREM 3:** *If  $0 < \lambda < 1$ ,  $\gamma(\lambda) = 1$ , and  $P_\lambda^*$  is given by (A10) in Appendix B, then  $(\lambda, P_\lambda^*)$  is an overall equilibrium. If  $\gamma(1) < 1$ , then  $(1, P_1^*)$  is an overall equilibrium. If  $\gamma(0) > 1$ , then  $(0, P_0^*)$  is an overall equilibrium. For all price equilibria  $P_\lambda$  which are monotone functions of  $w_\lambda$ , there exists a unique overall equilibrium  $(\lambda, P_\lambda^*)$ .*

**PROOF:**

The first three sentences follow immediately from the definition of overall equilibrium given above equation (12), and Theorems 1 and 2. Uniqueness follows from the monotonicity of  $\gamma(\cdot)$  which follows from (A11) and (14). The last two sentences in the statement of the theorem follow immediately.

In the process of proving Theorem 3, we have noted

**COROLLARY 1:**  $\gamma(\lambda)$  is a strictly monotone increasing function of  $\lambda$ .

This looks paradoxical; we expect the ratio of informed to uninformed expected utility to be a decreasing function of  $\lambda$ . But, we have defined utility as negative. Therefore



as  $\lambda$  rises, the expected utility of informed traders does go down relative to uninformed traders.

Note that the function  $\gamma(0) = e^{ac}(\text{Var}(u^*|\theta))/\text{Var } u^*)^{1/2}$ . Figure 1 illustrates the determination of the equilibrium  $\lambda$ . The figure assumes that  $\gamma(0) < 1 < \gamma(1)$ .

### G. Characterization of Equilibrium

We wish to provide some further characterization of the equilibrium. Let us define

$$(16a) \quad m = \left( \frac{a\sigma_\theta^2}{\lambda} \right)^2 \frac{\sigma_x^2}{\sigma_\theta^2}$$

$$(16b) \quad n = \frac{\sigma_\theta^2}{\sigma_\epsilon^2}$$

Note that  $m$  is inversely related to the informativeness of the price system since the squared correlation coefficient between  $P_\lambda^*$  and  $\theta^*$ ,  $\rho_\theta^2$  is given by

$$(17) \quad \rho_\theta^2 = \frac{1}{1+m}$$

Similarly,  $n$  is directly related to the quality of the informed trader's information because  $n/(1+n)$  is the squared correlation coefficient between  $\theta^*$  and  $u^*$ .

Equations (14) and (15) show that the cost of information  $c$ , determines the equilibrium ratio of information quality between informed and uninformed traders ( $\text{Var}(u^*|\theta)/\text{Var}(u^*|w_\lambda)$ ). From (1), (A11) of Appendix A, and (16), this can be written as

$$(18) \quad \frac{\text{Var}(u^*|\theta)}{\text{Var}(u^*|w_\lambda)} = \frac{1+m}{1+m+nm} = \left( 1 + \frac{nm}{1+m} \right)^{-1}$$

Substituting (18) into (14) and using (15) we obtain, for  $0 < \lambda < 1$ , in equilibrium

$$(19a) \quad m = \frac{e^{2ac} - 1}{1 + n - e^{2ac}}$$

or

$$(19b) \quad 1 - \rho_\theta^2 = \frac{e^{2ac} - 1}{n}$$

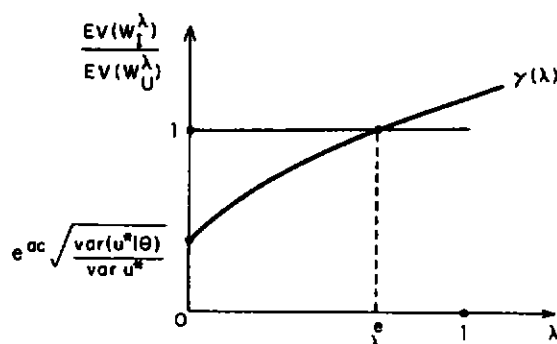


FIGURE 1

Note that (19) holds for  $\gamma(0) < 1 < \gamma(1)$ , since these conditions insure that the equilibrium  $\lambda$  is between zero and one. Equation (19b) shows that the equilibrium informativeness of the price system is determined completely by the cost of information  $c$ , the quality of the informed trader's information  $n$ , and the degree of risk aversion  $a$ .

### H. Comparative Statics

From equation (19b), we immediately obtain some basic comparative statics results:

1) An increase in the quality of information ( $n$ ) increases the informativeness of the price system.

2) A decrease in the cost of information increases the informativeness of the price system.

3) A decrease in risk aversion leads informed individuals to take larger positions, and this increases the informativeness of the price system.

Further, all other changes in parameters, such that  $n$ ,  $a$ , and  $c$  remain constant, do not change the equilibrium degree of informativeness of the price system; other changes lead only to particular changes in  $\lambda$  of a magnitude to exactly offset them. For example:

4) An increase in noise ( $\sigma_\epsilon^2$ ) increases the proportion of informed traders. At any given  $\lambda$ , an increase in noise reduces the informativeness of the price system; but it increases the returns to information and leads more individuals to become informed; the remarkable result obtained above establishes that the two effects exactly offset each

other so that the equilibrium informativeness of the price system is unchanged. This can be illustrated diagrammatically if we note from (16a) that for a given  $\lambda$ , an increase in  $\sigma_x^2$  raises  $m$  which from (18) lowers  $(\text{Var}(u^*|\theta))/\text{Var}(u^*|w_\lambda)$ . Thus from (14) a rise in  $\sigma_x^2$  leads to a vertical downward shift of the  $\gamma(\lambda)$  curve in Figure 1, and thus a higher value of  $\lambda^e$ .

5) Similarly an increase in  $\sigma_e^2$  for a constant  $n$  (equivalent to an increase in the variance of  $u$  since  $n$  is constant) leads to an increased proportion of individuals becoming informed—and indeed again just enough to offset the increased variance, so that the degree of informativeness of the price system remains unchanged. This can also be seen from Figure 1 if (16) is used to note that an increase in  $\sigma_e^2$  with  $n$  held constant by raising  $\sigma_\theta^2$  leads to an increase in  $m$  for a given  $\lambda$ . From (18) and (14) this leads to a vertical downward shift of the  $\gamma(\lambda)$  curve and thus a higher value of  $\lambda^e$ .

6) It is more difficult to determine what happens if, say  $\sigma_\theta^2$  increases, keeping  $\sigma_u^2$  constant (implying a fall in  $\sigma_e^2$ ), that is, the information obtained is more informative. This leads to an increase in  $n$ , which from (19b) implies that the equilibrium informativeness of the price system rises. From (16) it is clear that  $m$  and  $nm$  both fall when  $\sigma_\theta^2$  rises (keeping  $\sigma_u^2 = \sigma_\theta^2 + \sigma_e^2$  constant). This implies that the  $\gamma(\lambda)$  curve may shift up or down depending on the precise values of  $c$ ,  $a$ , and  $n$ .<sup>8</sup> This ambiguity arises because an

<sup>8</sup>From (14) and (18) it is clear that  $\lambda$  rises if and only if  $\text{Var}(u^*|\theta) + \text{Var}(u^*|w_\lambda)$  falls due to the rise in  $\sigma_\theta^2$  for a given  $\lambda$ . This occurs if and only if  $nm/(1+m)$  rises. Using (16) to differentiate  $nm/(1+m)$  with respect to  $\sigma_\theta^2$  subject to the constraint that  $d\sigma_u^2 = 0$  (i.e.,  $d\sigma_\theta^2 = -d\sigma_e^2$ ), we find that the sign of

$$\begin{aligned} \frac{d}{d\sigma_\theta^2} \left( \frac{nm}{1+m} \right) &= \text{sgn} \left[ m \left( \frac{n+1}{n} \right) - 1 \right] \\ &= \text{sgn} \left[ \left( \frac{\gamma}{n-\gamma} \right) \left( \frac{n+1}{n} \right) - 1 \right] \end{aligned}$$

where  $\gamma \equiv e^{2ac} - 1$  and the last equality follows from equation (19a). Thus for  $n$  very large the derivative is negative so that  $\lambda$  falls due to an increase in the precision of the informed trader's information. Similarly if  $n$  is sufficiently small, the derivative is positive and thus  $\lambda$  rises.

improvement in the precision of informed traders' information, with the cost of the information fixed, increases the benefit of being informed. However, some of the improved information is transmitted, via a more informative price system, to the uninformed; this increases the benefits of being uninformed. If  $n$  is small, both the price system  $m$  is not very informative and the marginal value of information to informed traders is high. Thus the relative benefits of being informed rises when  $n$  rises; implying that the equilibrium  $\lambda$  rises. Conversely when  $n$  is large the price system is very informative and the marginal value of information is low to informed traders so the relative benefits of being uninformed rises.

7) From (14) it is clear that an increase in the cost of information  $c$  shifts the  $\gamma(\lambda)$  curve up and thus decreases the percentage of informed traders.

The above results are summarized in the following theorem.

**THEOREM 4:** For equilibrium  $\lambda$  such that  $0 < \lambda < 1$ :

A. The equilibrium informativeness of the price system,  $\rho_\theta^2$ , rises if  $n$  rises,  $c$  falls, or  $a$  falls.

B. The equilibrium informativeness of the price system is unchanged if  $\sigma_x^2$  changes, or if  $\sigma_u^2$  changes with  $n$  fixed.

C. The equilibrium percentage of informed traders will rise if  $\sigma_x^2$  rises,  $\sigma_u^2$  rises for a fixed  $n$ , or  $c$  falls.

D. If  $\bar{n}$  satisfies  $(e^{2ac} - 1)/(\bar{n} - (e^{2ac} - 1)) = \bar{n}/(\bar{n} + 1)$ , then  $n >^{(<)} \bar{n}$  implies that  $\lambda$  falls (rises) due to an increase in  $n$ .

**PROOF:**

Parts A–C are proved in the above remarks. Part D is proved in footnote 8.

#### I. Price Cannot Fully Reflect Costly Information

We now consider certain limiting cases, for  $\gamma(0) < 1 < \gamma(1)$ , and show that equilibrium does not exist if  $c > 0$  and price is fully informative.

1) As the cost of information goes to zero, the price system becomes more infor-

mative, but at a positive value of  $c$ , say  $\hat{c}$ , all traders are informed. From (14) and (15)  $\hat{c}$  satisfies

$$e^{ac} \sqrt{\frac{\text{Var}(u^*|\theta)}{\text{Var}(u^*|w_1)}} = 1$$

2) From (19a) as the precision of the informed trader's information  $n$  goes to infinity, i.e.,  $\sigma_e^2 \rightarrow 0$  and  $\sigma_\theta^2 \rightarrow \sigma_u^2$ ,  $\sigma_u^2$  held fixed, the price system becomes perfectly informative. Moreover the percentage of informed traders goes to zero! This can be seen from (18) and (15). That is, as  $\sigma_e^2 \rightarrow 0$ ,  $nm/(1+m)$  must stay constant for equilibrium to be maintained. But from (19b) and (17),  $m$  falls as  $\sigma_e^2$  goes to zero. Therefore  $nm$  must fall, but  $nm$  must not go to zero or else  $nm/(1+m)$  would not be constant. From (16)  $nm = (a/\lambda)^2 \sigma_e^2 \sigma_x^2$ , and thus  $\lambda$  must go to zero to prevent  $nm$  from going to zero as  $\sigma_e^2 \rightarrow 0$ .

3) From (16a) and (19a) it is clear that as noise  $\sigma_x^2$  goes to zero, the percentage of informed traders goes to zero. Further, since (19a) implies that  $m$  does not change as  $\sigma_x^2$  changes, the informativeness of the price system is unchanged as  $\sigma_x^2 \rightarrow 0$ .

Assume that  $c$  is small enough so that it is worthwhile for a trader to become informed when no other trader is informed. Then if  $\sigma_x^2 = 0$  or  $\sigma_e^2 = 0$ , there exists no competitive equilibrium. To see this, note that equilibrium requires either that the ratio of expected utility of the informed to the uninformed be equal to unity, or that if the ratio is larger than unity, no one be informed. We shall show that when no one is informed, it is less than unity so that  $\lambda = 0$  cannot be an equilibrium; but when  $\lambda > 0$ , it is greater than unity. That is, if  $\sigma_x^2 = 0$  or  $\sigma_e^2 = 0$ , the ratio of expected utilities is not a continuous function of  $\lambda$  at  $\lambda = 0$ .

This follows immediately from observing that at  $\lambda = 0$ ,  $\text{Var}(u^*|w_0) = \text{Var} u^*$ , and thus by (14)

$$\begin{aligned} (20) \quad \frac{EV(W_{li}^0)}{EV(W_{li}^0)} &= e^{ac} \sqrt{\frac{\sigma_e^2}{\sigma_e^2 + \sigma_\theta^2}} \\ &= e^{ac} \sqrt{\frac{1}{1+n}} \end{aligned}$$

while if  $\lambda > 0$ , by (18)

$$\frac{EV(W_{li}^\lambda)}{EV(W_{li}^\lambda)} = e^{ac} \sqrt{\frac{1}{1+n \frac{m}{m+1}}}$$

But if  $\sigma_x^2 = 0$  or  $\sigma_e^2 = 0$ , then  $m = 0$ ,  $nm = 0$  for  $\lambda > 0$ , and hence

$$(21) \quad \lim_{\lambda \rightarrow 0} \frac{EV(W_{li}^\lambda)}{EV(W_{li}^\lambda)} = e^{ac}$$

It immediately follows that

**THEOREM 5:** (a) *If there is no noise ( $\sigma_x^2 = 0$ ), an overall equilibrium does not exist if (and only if)  $e^{ac} < \sqrt{1+n}$ .* (b) *If information is perfect ( $\sigma_e^2 = 0, n = \infty$ ), there never exists an equilibrium.*

**PROOF:**

(a) If  $e^{ac} < \sqrt{1+n}$ , then by (20) and (21),  $\gamma(\lambda)$  is discontinuous at  $\lambda = 0$ ;  $\lambda = 0$  is not an equilibrium since by (20)  $\gamma(0) < 1$ ;  $\lambda > 0$  is not an equilibrium since by (21)  $\gamma(\lambda) > 1$ .

(b) If  $\sigma_e^2 = 0$  and  $\sigma_\theta^2 = \sigma_u^2$  so that information is perfect, then for  $\lambda > 0$ ,  $nm = 0$  by (16) and hence  $\gamma(\lambda) > 1$  by (21). From (20)  $\gamma(0) = 0 < 1$ .

If there is no noise and some traders become informed, then *all* their information is transmitted to the uninformed by the price system. Hence each informed trader acting as a price taker thinks the informativeness of the price system will be unchanged if he becomes uninformed, so  $\lambda > 0$  is not an equilibrium. On the other hand, if no traders are informed, then each uninformed trader learns nothing from the price system, and thus he has a desire to become informed (if  $e^{ac} < (1+n)^{1/2}$ ). Similarly if the informed traders get perfect information, then their demands are very sensitive to their information, so that the market-clearing price becomes very sensitive to their information and thus reveals  $\theta$  to the uninformed. Hence all traders desire to be uninformed. But if all traders are uninformed, each trader can eliminate the risk of his portfolio by the purchase of information, so each trader desires to be informed.

In the next section we show that the non-existence of competitive equilibrium can be thought of as the breakdown of competitive markets due to lack of trade. That is, we will show that as  $\sigma_x^2$  gets very small, trade goes to zero and markets serve no function. Thus competitive markets close for lack of trade "before" equilibrium ceases to exist at  $\sigma_x^2 = 0$ .

### III. On the Thinness of Speculative Markets

In general, trade takes place because traders differ in endowments, preferences, or beliefs. Grossman (1975, 1977, 1978) has argued that differences in preferences are not a major factor in explaining the magnitude of trade in speculative markets. For this reason the model in Section II gave all traders the same risk preferences (note that none of the results in Section II are affected by letting traders have different coefficients of absolute risk aversion). In this section we assume that trade requires differences in endowments or beliefs and dispense with differences in risk preference as an explanatory variable.<sup>9</sup>

There is clearly some fixed cost in operating a competitive market. If traders have to bear this cost, then trade in the market must be beneficial. Suppose traders have the same endowments and beliefs. Competitive equilibrium will leave them with allocations which are identical with their initial endowments. Hence, if it is costly to enter such a competitive market, no trader would ever enter. We will show below that in an important class of situations, there is continuity in the amount of net trade. That is, when initial endowments are the same and peo-

ples' beliefs differ *slightly*, then the competitive equilibrium allocation that an individual gets will be only *slightly* different from his initial endowment. Hence, there will only be a slight benefit to entering the competitive market. This could, for sufficiently high operating costs, be outweighed by the cost of entering the market.

The amount of trade occurring at any date is a random variable; a function of  $\theta$  and  $x$ . It is easy to show that it is a normally distributed random variable. Since one of the primary determinants of the size of markets is differences in beliefs, one might have conjectured that markets will be thin, in some sense, if almost all traders are either informed or uninformed. This is not, however, obvious, since the amount of trade by any single trader may be a function of  $\lambda$  as well, and a few active traders can do the job of many small traders. In our model, there is a sense, however, in which our conjecture is correct.

We first calculate the magnitude of trades as a function of the exogenous parameters,  $\theta$  and  $x$ . Let  $h \equiv \sigma_\epsilon^2$ ,  $\bar{x} = Ex^*$ , and  $\bar{\theta} \equiv E\theta^*$ . (The actual trades will depend on the distribution of random endowments across all of the traders, but these we shall net out.) Per capita net trade is<sup>10</sup>

$$(22) \quad X_I - x = (1-\lambda) \left[ \left( nm + \frac{ah}{\lambda} \right) (x - \bar{x}) + [(m+1)n - 1](\theta - \bar{\theta}) + \bar{x}nm \right] + [1 + m + \lambda nm]$$

<sup>10</sup>Calculation of distribution of net trades

$$\begin{aligned} & \frac{\lambda}{ah} (\theta - RP_\lambda) \\ & + \frac{(1-\lambda) \left[ (\bar{\theta} - RP_\lambda)(1+m)n + \theta - \bar{\theta} - \frac{ah}{\lambda} (x - \bar{x}) \right]}{ah(1+m+nm)n} = x \\ & \text{or } \frac{(\theta - RP_\lambda)}{ah} \left( \lambda + \frac{(1-\lambda)(1+m)}{1+m+nm} \right) \\ & = \left( \frac{\theta - RP_\lambda}{ah} \right) \left( \frac{1+m+\lambda nm}{1+m+nm} \right) \\ & = x + \frac{(1-\lambda) \left[ [(m+1)n - 1](\theta - \bar{\theta}) + \frac{ah}{\lambda} (x - \bar{x}) \right]}{ah(1+m+\lambda nm)n} \end{aligned}$$

<sup>9</sup>In the model described in Section II it was assumed that an individual's endowment  $\bar{X}_i$  is independent of the market's per capita endowment  $x^*$ . This was done primarily so there would not be useful information in an individual's endowment about the total market endowment. Such information would be useful in equilibrium because an individual observes  $P_\lambda(\theta, x)$ . If due to observing  $\bar{X}_i$ , he knows something about  $x$ , then by observing  $P_\lambda(\theta, x)$ ,  $\bar{X}_i$  is valuable in making inferences about  $\theta$ . To take this into account is possible, but would add undue complication to a model already overburdened with computations.

Thus, the mean of total informed trade is

$$(23) \quad E\lambda(X_I - x) = \frac{(1-\lambda)\lambda m \bar{x}}{1+m+\lambda nm}$$

and its variance is

$$(24) \quad \sigma_\theta^2(1-\lambda)^2\lambda^2 \left[ \left[ (m+1)n-1 \right]^2 + \left( nm + \frac{a\sigma_\epsilon^2}{\lambda} \right)^2 \frac{\sigma_x^2}{\sigma_\theta^2} \right] + (1+m+\lambda nm)^2 n^2$$

In the last section we considered limiting values of the exogenous variables with the property that  $\lambda \rightarrow 0$ . The following theorem will show that the mean and variance of trade go to zero as  $\lambda \rightarrow 0$ . That is, the distribution of  $\lambda(X_I - x)$  becomes degenerate at zero as  $\lambda \rightarrow 0$ . This is not trivial because as  $\lambda \rightarrow 0$  due to  $n \rightarrow \infty$  (very precise information), the informed trader's demand  $X_I(P, \theta)$  goes to infinity at most prices because the risky asset becomes riskless with perfect information.

**THEOREM 6:** (a) *For sufficiently large or small  $c$ , the mean and variance of trade is zero.* (b) *As the precision of informed traders' information  $n$  goes to infinity, the mean and variance of trade go to zero.*

**PROOF:**

(a) From remark 1) in Section II, Part I,  $\lambda = 1$  if  $c \leq \hat{c}$ , which from (23) and (24) implies trade is degenerate at zero. From (14), for  $c$  sufficiently large, say  $c^0$ ,  $\gamma(0) = 1$ , so

$$\begin{aligned} \text{or } X_I &= \frac{1+m+nm}{1+m+\lambda nm} \\ &\times \left[ x + \frac{(1-\lambda) \left[ [(m+1)-1](\theta - \bar{\theta}) + \frac{ah}{\lambda}(x - \bar{x}) \right]}{ah(1+m+nm)n} \right] \\ X_I - x &= \\ &\frac{(1-\lambda) \left[ \left( nm + \frac{ah}{\lambda} \right) (x - \bar{x}) + [(m+1)-1](\theta - \bar{\theta}) + \bar{x}nm \right]}{(1+m+\lambda nm)n} \end{aligned}$$

the equilibrium  $\lambda = 0$ . As  $c$  goes to  $c^0$  from below  $\lambda \rightarrow 0$ , and from (14), (15), and (18)  $\lim_{c \uparrow c^0} (1 + nm/(1+m))^{-1/2} = e^{-ac^0}$ . Hence  $\lim_{c \uparrow c^0} (nm/(1+m))$  is a finite positive number. Thus from (22) mean trade goes to zero as  $c \uparrow c^0$ . If the numerator and the denominator of (24) are divided by  $(1+m)^2$ , then again using the fact that  $m/(1+m)$  has a finite limit gives the result that as  $c \uparrow c^0$ ,  $\lambda \rightarrow 0$ , and variance of trade goes to zero.

(b) By (14), (15), and (18),  $nm/(1+m)$  is constant as  $n \rightarrow \infty$ . Further, from remark 2) of Section II, Part I,  $\lambda \rightarrow 0$  as  $n \rightarrow \infty$ . Hence from (23) and (24), the mean and variance of trade go to zero.

(c) From remark 3) in Section II, Part I,  $m$  is constant and  $\lambda$  goes to zero as  $\sigma_x^2 \rightarrow 0$ . Therefore mean trade goes to zero. In (24), note that  $(nm + a\sigma_\epsilon^2/\lambda)^2 \sigma_x^2 / \sigma_\theta^2 = (nm\sigma_x/\sigma_\theta + (m)^{1/2})^2$  by (16a). Hence the variance of trade goes to zero as  $\sigma_x^2 \rightarrow 0$ .

Note further that  $\lambda(X_I - x) + (1-\lambda)(X_U - x) = 0$  implies that no trade will take place as  $\lambda \rightarrow 1$ . Thus, the result that competitive equilibrium is incompatible with informationally efficient markets should be interpreted as meaning that speculative markets where prices reveal a lot of information will be very thin because it will be composed of individuals with very similar beliefs.

#### IV. On the Possibility of Perfect Markets

In Section II we showed that the price system reveals the signal  $w_\lambda^*$  to traders, where

$$w_\lambda \equiv \theta - \frac{a\sigma_\epsilon^2}{\lambda}(x - Ex^*)$$

Thus, for given information of informed traders  $\theta$ , the price system reveals a noisy version of  $\theta$ . The noise is  $(a\sigma_\epsilon^2/\lambda)(x - Ex^*)$ . Uninformed traders learn  $\theta$  to within a random variable with mean zero and variance  $(a\sigma_\epsilon^2/\lambda)^2 \text{Var } x^*$ , where  $\sigma_\epsilon^2$  is the precision of informed traders' information,  $\text{Var } x^*$  is the amount of endowment uncertainty,  $\lambda$  the fraction of informed traders, and  $a$  is the degree of absolute risk aversion. Thus, in general the price system does not reveal all



the information about "the true value" of the risky asset. ( $\theta$  is the true value of the risky asset in that it reflects the best available information about the asset's worth.)

The only way informed traders can earn a return on their activity of information gathering, is if they can use their information to take positions in the market which are "better" than the positions of uninformed traders. "Efficient Markets" theorists have claimed that "at any time prices fully reflect all available information" (see Eugene Fama, p. 383). If this were so then informed traders could not earn a return on their information.

We showed that when the efficient markets hypothesis is true and information is costly, competitive markets break down. This is because when  $\sigma_e^2 = 0$  or  $Var x^* = 0$ ,  $w_\lambda$ , and thus price, does reflect all the information. When this happens, each informed trader, because he is in a competitive market, feels that he could stop paying for information and do as well as a trader who pays nothing for information. But all informed traders feel this way. Hence having any positive fraction informed is not an equilibrium. Having no one informed is also not an equilibrium, because then each trader, taking the price as given, feels that there are profits to be made from becoming informed.

Efficient Markets theorists seem to be aware that costless information is a *sufficient* condition for prices to fully reflect all available information (see Fama, p. 387); they are not aware that it is a *necessary* condition. But this is a *reductio ad absurdum*, since price systems and competitive markets are important only when information is costly (see Fredrick Hayek, p. 452).

We are attempting to redefine the Efficient Markets notion, not destroy it. We have shown that when information is very inexpensive, or when informed traders get very precise information, then equilibrium exists and the market price will reveal most of the informed traders' information. However, it was argued in Section III that such markets are likely to be thin because traders have almost homogeneous beliefs.

There is a further conflict. As Grossman (1975, 1977) showed, whenever there are differences in beliefs that are not completely arbitrated, there is an incentive to create a market. (Grossman, 1977, analyzed a model of a storable commodity whose spot price did not reveal all information because of the presence of noise. Thus traders were left with differences in beliefs about the future price of the commodity. This led to the opening of a futures market. But then uninformed traders had two prices revealing information to them, implying the elimination of noise.) But, because differences in beliefs are themselves endogenous, arising out of expenditure on information and the informativeness of the price system, the creation of markets eliminates the differences of beliefs which gave rise to them, and thus causes those markets to disappear. If the creation of markets were costless, as is conventionally assumed in equilibrium analyses, equilibrium would never exist. For instance, in our model, were we to introduce an additional security, say a security which paid

$$z = \begin{cases} 1 & \text{if } u > E\theta^* \\ 0 & \text{if } u \leq E\theta^* \end{cases}$$

then the demand  $y$  for this security by the informed would depend on its price, say  $q$  on  $p$  and on  $\theta$ , while the uninformed demand depends only on  $p$  and  $q$ :

$$\lambda y_I(q, p, \theta) + (1 - \lambda) y_U(q, p) = 0$$

is the condition that demand equals (supply is zero for a pure security). Under weak assumptions,  $q$  and  $p$  would convey all the information concerning  $\theta$ . Thus, the market would be "noiseless" and no equilibrium could exist.

Thus, we could argue as soon as the assumptions of the conventional perfect capital markets model are modified to allow even a slight amount of information imperfection and a slight cost of information, the traditional theory becomes untenable. There *cannot* be as many securities as states of nature. For if there were, competitive equilibrium would not exist.

It is only because of costly transactions and the fact that this leads to there being a limited number of markets, that competitive equilibrium can be established.

We have argued that because information is costly, prices cannot perfectly reflect the information which is available, since if it did, those who spent resources to obtain it would receive no compensation. There is a fundamental conflict between the efficiency with which markets spread information and the incentives to acquire information. However, we have said nothing regarding the social benefits of information, nor whether it is socially optimal to have "informationally efficient markets." We hope to examine the welfare properties of the equilibrium allocations herein in future work.

## APPENDIX A

Here we collect some facts on conditional expectations used in the text. If  $X^*$  and  $Y^*$  are jointly normally distributed then

$$(A1) \quad E[X^*|Y^*=Y] \\ = EX^* + \frac{\text{Cov}(X^*, Y^*)}{\text{Var}(Y^*)} (Y - EY^*)$$

$$(A2) \quad \text{Var}[X^*|Y^*=Y] \\ = \text{Var}(X^*) - \frac{[\text{Cov}(X^*, Y^*)]^2}{\text{Var}(Y^*)}$$

(See Paul Hoel, p. 200.) From (A1) note that  $E[X^*|Y^*]$  is a function of  $Y$ . If the expectation of both sides of (A1) is taken, we see that

$$(A3) \quad E\{E[X^*|Y^*=Y]\} = EX^*$$

Note that  $\text{Var}[X^*|Y^*=Y]$  is not a function of  $Y$ , as  $\text{Var}(X^*)$ ,  $\text{Cov}(X^*, Y^*)$ , and  $\text{Var}(Y^*)$  are just parameters of the joint distribution of  $X^*$  and  $Y^*$ .

Two other relevant properties of conditional expectation are

$$(A4) \quad E\{E[Y^*|F(X^*)]|X^*\} = E[Y^*|F(X^*)]$$

$$(A5) \quad E\{E[Y^*|X]|F(X^*)\} = E[Y^*|F(X^*)]$$

where  $F(\cdot)$  is a given function on the range of  $X^*$  (see Robert Ash, p. 260).

## APPENDIX B

PROOF of Theorem 1:

(a) Suppose  $\lambda = 0$ ; then (9) becomes

$$(A6) \quad X_U(P_0(\theta, x), P_0^*) = x$$

Define

$$(A7) \quad P_0(\theta, x) \equiv \frac{E\theta^* - a\sigma_u^2}{R}$$

where  $\sigma_u^2$  is the variance of  $u$ . Note that  $P_0(\theta^*, x^*)$  is uncorrelated with  $u^*$ , as  $x^*$  is uncorrelated with  $u^*$ . Hence

$$(A8) \quad E[u^*|P_0^* = P_0(\theta, x)] = Eu^* = E\theta^*$$

$$\text{and } \text{Var}[u^*|P_0^* = P_0(\theta, x)] = \text{Var}[u^*]$$

Substitution of (A8) in (8) yields

$$(A9) \quad X_U(P_0^*, P_0(\theta, x)) = \frac{E\theta^* - RP_0(\theta, x)}{a \text{Var} u}$$

Substitution of (A7) in the right-hand side of (A9) yields  $X_U(P_0^*(\theta, x), P_0^*) = x$  which was to be shown.

(b) Suppose  $0 < \lambda < 1$ . Let

$$(A10) \quad P_\lambda(\theta, x) = \frac{\frac{\lambda w_\lambda}{a\sigma_\epsilon^2} + \frac{(1-\lambda)E[u^*|w_\lambda]}{a \text{Var}[u^*|w_\lambda]} - Ex^*}{R \left[ \frac{\lambda}{a\sigma_\epsilon^2} + \frac{(1-\lambda)}{a \text{Var}[u^*|w_\lambda]} \right]}$$

Note that from equations (1), (10), (A1) and (A2):

(A11a)

$$E(u^*|w_\lambda) = E\theta^* + \frac{\sigma_\theta^2}{\text{Var} w_\lambda} (w_\lambda - E\theta^*)$$

$$(A11b) \quad \text{Var}(u^*|w_\lambda) = \sigma_\theta^2 + \sigma_\epsilon^2 - \frac{\sigma_\theta^2}{\text{Var} w_\lambda}$$

$$(A11c) \quad \text{Var} w_\lambda = \sigma_\theta^2 + \left( \frac{a\sigma_\epsilon^2}{\lambda} \right)^2 \text{Var} x^*$$

Since  $P_\lambda(\theta, x)$  is a linear function of  $w_\lambda$ , it is immediate that  $E(u^*|w_\lambda) \equiv E(u^*|P_\lambda)$ ,  $Var(u^*|w_\lambda) = Var(u^*|P_\lambda)$ , etc. To see that  $P_\lambda^*$  is an equilibrium, we must show that the following equation holds as an identity in  $(\theta, x)$ , for  $P_\lambda(\cdot)$  defined by (A10):

$$(A12) \quad \lambda \cdot \frac{\theta - RP_\lambda}{a\sigma_e^2} + (1-\lambda) \frac{E[u^*|w_\lambda] - RP_\lambda}{a Var[u^*|w_\lambda]} = x$$

It is immediate from (10) that (A12) holds as an identity in  $\theta$  and  $x$ .

PROOF of Theorem 2:

(a) *Calculation of the expected utility of the informed.* Using the fact that  $W_{ii}^\lambda$  is normally distributed conditional on  $(\bar{X}_i, \theta, x)$

$$(A13) \quad E[V(W_{ii}^\lambda)|\bar{X}_i, \theta, x] \\ = \exp \left[ -a \left\{ E[W_{ii}^\lambda|\bar{X}_i, \theta, x] - \frac{a}{2} Var[W_{ii}^\lambda|\bar{X}_i, \theta, x] \right\} \right]$$

Using (8), (12), and the fact that  $(\theta, x)$  determines a particular  $P$ ,

$$(A14a) \quad E[W_{ii}^\lambda|\bar{X}_i, \theta, x] = R(W_{oi} - c) + \frac{(E[u^*|\theta] - RP_\lambda)^2}{a\sigma_e^2}$$

$$(A14b) \quad Var[W_{ii}^\lambda|\bar{X}_i, \theta, x] = \frac{(E[u^*|\theta] - RP_\lambda)^2}{a^2\sigma_e^2}$$

Substitution of (A14) into (A13) yields

$$(A15) \quad E[V(W_{ii}^\lambda)|\bar{X}_i, \theta, x] \\ = -\exp \left[ -aR(W_{oi} - c) - \frac{1}{2\sigma_e^2} (E[u^*|\theta] - RP_\lambda)^2 \right]$$

Note that, as  $P_\lambda^*(\cdot) = P_\lambda(\theta, x)$ ,

$$(A16) \quad E \left( E[V(W_{ii}^\lambda)|\bar{X}_i, \theta, x] | P_\lambda, \bar{X}_i \right) \\ = E[V(W_{ii}^\lambda)|P_\lambda, \bar{X}_i]$$

(see (A5)). Note that since  $W_{oi}$  is nonstochastic conditional on  $(P_\lambda, \bar{X}_i)$ , equation (A15) implies

$$(A17) \quad E[V(W_{ii}^\lambda)|P_\lambda, \bar{X}_i] = -\exp[-aR(W_{oi}^\lambda - c)] \cdot E \left[ \exp \left[ -\frac{1}{2\sigma_e^2} (E[u|\theta] - RP_\lambda)^2 \right] | P_\lambda, \bar{X}_i \right]$$

Note that by Theorem 1, conditioning on  $w_\lambda^*$  is equivalent to conditioning on  $P_\lambda^*$ . Define

$$(A18) \quad h_\lambda \equiv Var(E[u^*|\theta]|w_\lambda) \\ = Var(\theta|w_\lambda), h_0 \equiv \sigma_e^2 \equiv h$$

$$(A19) \quad Z \equiv \frac{E[u^*|\theta] - RP_\lambda}{\sqrt{h_\lambda}}$$

Using (3) and (A18), equation (A17) can be written as

$$(A20) \quad E[V(W_{ii}^\lambda)|P_\lambda, \bar{X}_i] \\ = e^{ac} V(RW_{oi}) E \left[ \exp \left[ -\frac{h_\lambda}{2\sigma_e^2} Z^2 \right] | w_\lambda \right]$$

since  $\bar{X}_i$  and  $w_\lambda$  are independent. Conditional on  $w_\lambda$ ,  $P_\lambda$  is nonstochastic and  $E[u^*|\theta]$  is normal. Hence conditional on  $w_\lambda$ ,  $(Z^*)^2$  has a noncentral *chi*-square distribution (see C. Rao, p. 181). Then for  $t > 0$  the moment generating function for  $(Z^*)^2$  can be written

$$(A21) \quad E[e^{-tZ^2}|w_\lambda] \\ = \frac{1}{\sqrt{1+2t}} \exp \left[ \frac{-(E[Z|w_\lambda])^2 t}{1+2t} \right]$$



Note that  $E[u^*|\theta] = E[u^*|\theta, x]$ . Hence

$$(A22) \quad E[E[u^*|\theta]|w_\lambda] = E[u^*|w_\lambda] \\ = E\theta^* + \frac{\sigma_\theta^2}{\text{Var } w_\lambda} (w_\lambda - E\theta^*)$$

since  $w_\lambda$  is just a function of  $(\theta, x)$ . Therefore

$$(A23) \quad E[Z^*|w_\lambda] = \frac{E[u^*|w_\lambda] - RP_\lambda}{\sqrt{h_\lambda}}$$

Since  $u = \theta + \varepsilon$

$$(A24) \quad \text{Var}(u^*|w_\lambda) = \sigma_\varepsilon^2 + \text{Var}(\theta^*|w_\lambda) = \sigma_\varepsilon^2 + h_\lambda$$

The nondegeneracy assumptions on  $(x^*, \varepsilon^*, u^*)$  imply  $h_\lambda > 0$ . Set  $t = (h_\lambda/2\sigma_\varepsilon^2)$ ; and evaluate (A21) using (A23) and (A24):

$$(A25) \quad E\left[\exp\left[-\frac{h_\lambda}{2\sigma_\varepsilon^2} Z^2\right] | w_\lambda\right] = \sqrt{\frac{\text{Var}(u^*|\theta)}{\text{Var}(u^*|w_\lambda)}} \\ \cdot \exp\left(\frac{-(E(u^*|w_\lambda) - RP_\lambda)^2}{2 \text{Var}(u^*|w_\lambda)}\right)$$

This permits the evaluation of (A20).

(b) Calculation of expected utility of the uninformed. Equations (8), (5), and the normality of  $W_{Ui}^\lambda$  conditional on  $w_\lambda$  can be used to show, by calculations parallel to (A13)–(A25), that

$$(A26) \quad E[V(W_{Ui}^\lambda)|w_\lambda, \bar{X}_i] \\ = V(RW_{0i}) \exp\left(\frac{-(E(u^*|w_\lambda) - RP_\lambda)^2}{2 \text{Var}(u^*|w_\lambda)}\right)$$

Hence

$$(A27) \quad E[V(W_{Ui}^\lambda)|w_\lambda, \bar{X}_i] - E[V(W_{Ui}^\lambda)|w_\lambda, \bar{X}_i] \\ = \left[ e^{ac} \sqrt{\frac{\text{Var}(u^*|\theta)}{\text{Var}(u^*|w_\lambda)}} - 1 \right] \\ \times E[V(W_{Ui}^\lambda)|w_\lambda, \bar{X}_i]$$

Taking expectations of both sides of (A27) yields:

$$(A28) \quad E[V(W_{Ui}^\lambda)] - E[V(W_{Ui}^\lambda)] \\ = \left[ e^{ac} \sqrt{\frac{\text{Var}(u^*|\theta)}{\text{Var}(u^*|w_\lambda)}} - 1 \right] EV(W_{Ui}^\lambda)$$

Equation (13) follows immediately from (A28).

## REFERENCES

- Robert B. Ash, *Real Analysis and Probability*, New York 1972.
- E. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *J. Finance*, May 1970, 25, 383–417.
- J. R. Green, "Information, Efficiency and Equilibrium," disc. paper no. 284, Harvard Inst. Econ. Res., Mar. 1973.
- , "The Non-Existence of Informational Equilibria," *Rev. Econ. Stud.*, Oct. 1977, 44, 451–64.
- S. Grossman, "Essays on Rational Expectations," unpublished doctoral dissertation, Univ. Chicago 1975.
- , "On the Efficiency of Competitive Stock Markets Where Traders Have Diverse Information," *J. Finance*, May 1976, 31, 573–85.
- , "The Existence of Futures Markets, Noisy Rational Expectations and Informational Externalities," *Rev. Econ. Stud.*, Oct. 1977, 64, 431–49.
- , "Further Results on the Informational Efficiency of Competitive Stock Markets," *J. Econ. Theory*, June 1978, 18, 81–101.
- , R. Kihlstrom, and L. Mirman, "A Bayesian Approach to the Production of Information and Learning by Doing," *Rev. Econ. Stud.*, Oct. 1977, 64, 533–47.
- F. H. Hayek, "The Use of Knowledge in Society," *Amer. Econ. Rev.*, Sept. 1945, 35, 519–30.
- Paul G. Hoel, *Introduction to Mathematical Statistics*, New York 1962.
- R. Kihlstrom and L. Mirman, "Information and Market Equilibrium," *Bell. J. Econ.*, Spring 1975, 6, 357–76.

R. E. Lucas, Jr., "Expectations and the Neutrality of Money," *J. Econ. Theory*, Apr. 1972, 4, 103-24.

C. Rao, *Linear Statistical Inference and Its Applications*, New York 1965.

J. E. Stiglitz, "Perfect and Imperfect Capital Markets," paper presented to the Econometric Society, New Orleans 1971.

\_\_\_\_\_, "Information and Capital Markets," mimeo., Oxford Univ. 1974.

# Exhibit 55

12 January 2010

# Testing for Materiality in Volatile Markets

By **Dr. Branko Jovanovic** and **Edward Fox**

In securities litigation, cases sometimes hinge on whether information that was omitted or misstated was important to a reasonable investor. Financial experts are often then called upon to examine whether a news announcement that corrected information that was previously misrepresented had a material effect on the price of a company's stock. To address the issue of materiality, experts and courts often rely on event studies.<sup>1</sup> The results of these event studies are used to calculate inflation (the difference between the actual price and the price the stock would have traded at had there been no misrepresentations), which is then used to estimate damages suffered by purchasers of the stock. Inferences from an event study will be stronger when the period used to calculate the expected returns and expected volatility (the estimation period) and the period in which the alleged disclosure occurred (the disclosure period) are similar except for the release of news related to the alleged fraud.<sup>2</sup> In this paper, we show that the increase in overall market volatility in the wake of the 2008 credit crisis can cause the traditional event study methodology to be inaccurate. In particular, we show that use of the traditional methodology during a period of generally increased volatility may improperly yield a finding that an immaterial disclosure is statistically significant (i.e., a "false positive"). We propose several ways to increase the accuracy of an event study in periods of increased market volatility.

## Event study primer

A typical allegation in a securities shareholder class action is that a company failed to disclose negative information relevant to its investors. Plaintiffs will often allege that this information was made public at some point (the disclosure date) and caused the stock price drop. An event study is typically used to evaluate plaintiffs' claims.<sup>3</sup> The purpose of an event study is to measure the price movement of a security in response to new information. An event study is conceptually performed in two stages. First, a market model is created that predicts the returns of a stock based on the returns of a market index.<sup>4</sup> The market model separates the stock's returns into two parts: the portion of returns explained by the market index and the part attributable to company-specific factors. This latter portion, known as

the residual or abnormal return, includes any part of the return caused by factors unrelated to the general market movement such as firm-specific information released that day. The market model also provides a measure of the variability of the company-specific portion of the stock's returns, which is known as the standard error. Experts use this measure to assess the statistical significance of the price movement following a disclosure. Because statistical significance is a way of characterizing how unusual a result is, the more volatile the stock returns are estimated to be, the larger any price movement would have to be for it to be deemed statistically significant.<sup>5</sup>

In shareholder class actions, market models are often estimated over the year prior to the beginning of the class period, and thus do not directly measure the variability of the stock's returns during the disclosure period. Insofar as a stock's returns behave differently around the time of the disclosures than during the estimation period, the standard error of the market model may be an inaccurate measure of company-specific variability at the time of the disclosures. For cases with relatively brief class periods that coincide with stable securities markets, the assumption that the volatility has not changed substantially between the estimation and disclosure periods is more likely to be correct. For a case with a long class period that encompasses periods of market stability and periods of uncertain and tumultuous markets, traditional methods may have to be reconsidered.

### Recent years are marked by tumultuous markets

The period from mid-2008 to mid-2009 was one of the most volatile times in the modern history of the United States' stock market. The increase in volatility in the stock market is illustrated in Table 1. The average expected daily volatility (the average expected fluctuation in the returns) of the S&P 100 Index, remained stable from approximately 0.7% to 1% from mid-2003 to mid-2007. From mid-2007 to mid-2008, the volatility increased to 1.3%. Finally, from mid-2008 to mid-2009, the volatility reached 2.22%, almost triple the level during the relatively stable mid-2005 to mid-2007 period.

**Table 1**

<b>Period</b>	<b>Average Implied Volatility</b>
1 July 2003-30 June 2004	0.97%
1 July 2004-30 June 2005	0.79%
1 July 2005-30 June 2006	0.68%
1 July 2006-30 June 2007	0.68%
1 July 2007-30 June 2008	1.29%
1 July 2008-30 June 2009	2.22%

This increase in volatility was not confined to market, or "systemic," risk exhibited by the S&P 100. Firm-specific risk, as measured by the variability of company-specific returns, increased substantially among firms in the S&P 100. The effect was most pronounced in the financial sector, where the crisis started, but is evident among all firms. Table 2 summarizes results of an analysis conducted on the constituents of the S&P 100 Index, where the average company-specific volatility is calculated over all of the 100 companies which traded over the full period, as well as for a subset of the companies in financial and non-financial sectors.<sup>6</sup> The results are striking: the average volatility for the companies in the financial sector from mid-2008 to mid-2009 was over twice the average volatility as in the previous year, at more than 5.4%. This implies that even an average market-adjusted return in mid-2008 to mid-2009 for a financial company would be deemed statistically significant if compared to the average volatility in the mid-2007 to mid-2008 period.<sup>7</sup>

**Table 2**

Period	Average Company-Specific Volatility		
	All	Financial	Non-Financial
1 July 2005-30 June 2006	1.18%	0.84%	1.22%
1 July 2006-30 June 2007	1.14%	0.82%	1.19%
1 July 2007-30 June 2008	1.51%	1.88%	1.46%
1 July 2008-30 June 2009	2.69%	5.42%	2.30%

As a consequence of this increase in volatility, the traditional event study methodology can prove inaccurate. The recent increase in volatility calls into question whether inferences from an event study using an estimation period from a few years back can be accepted without additional analysis: an estimation period from mid-2005 to mid-2007 (or earlier) paired with a later event period may yield inaccurate and unreliable results, given the striking increase in firm-specific volatility.

### The traditional event study methodology yields too many “statistically significant” days

To determine how problematic this increase in market volatility is for traditional event studies, we analyzed the constituents of the S&P 100 Index to determine how many days would exhibit statistically significant abnormal returns during three one-year periods (1 July 2006-30 June 2007, 1 July 2007-30 June 2008, and 1 July 2008-30 June 2009) when using the prior year as the estimation period.<sup>8</sup> If the estimation period and the disclosure period are similar, using the standard 5% significance level, there is a 5% chance that a day will be deemed statistically significant in the absence of material news. In other words, we expect 5% of tested days to have statistically significant returns if the baseline market model is accurate.<sup>9</sup> If the standard errors are based on a period when the stock was much less volatile, they will be too low and too many days will be identified as statistically significant. If we estimated standard errors based on a period when the stock was much more volatile, we would find that too few days are found to be statistically significant.

Because about 5% of the approximately 252 trading days in a year should be statistically significant at the 5% level by definition, the expected number of significant days for each of the three periods is about 12 days. The average number of significant dates over 97 securities summarized in Table 3 confirms our concerns about the standard methodology: it performs well in the 1 July 2006-30 June 2007 period, which is expected given that the volatilities in that period and the prior year were roughly the same. In contrast, in the 1 July 2007-30 June 2008 period and particularly in the 1 July 2008-30 June 2009 period, the standard methodology yields far too many “statistically significant” days on average for each company.

**Table 3**

Period	Statistically Significant Dates	
	Number	Percent
1 July 2006-30 June 2007	11.7	4.7%
1 July 2007-30 June 2008	32.1	12.7%
1 July 2008-30 June 2009	52.6	20.9%

## Three possible solutions

### Move the estimation period forward to the event window

The simplest way to resolve the issues associated with performing an event study over a period of heightened volatility would be to use the disclosure period as the estimation period. One can “overlap” the estimation and the disclosure period, by construction guaranteeing similar volatilities between the two periods. The application of this approach, sometimes employed by plaintiffs’ experts, is not necessarily free from bias. One must ensure that all disclosures and misrepresentations, as well as any other news associated with the fraud, are excluded from the estimation period in order to obtain a “clean” benchmark. Determining relevancy introduces subjectivity. Further, when the pattern of company-specific returns is alleged to be affected by the fraud even on days not directly associated with the release of news, using the company-specific data from the disclosure period may be objectionable.<sup>10</sup>

### Obtain a market expectation of volatility for disclosure dates

Another possibility is to use the market’s expectation of daily volatility to measure statistical significance. Trading an option is essentially taking a bet on the volatility of the stock underlying it. Using the well-known Black-Scholes option-pricing formula, we can back out the market expectation of volatility from the market prices of traded options. An estimate of volatility based on a stock’s option price is called *implied volatility*. Since implied volatility may rise on the day of an alleged disclosure, reflecting a rise in uncertainty regarding the company’s prospects, implied volatility from the day before the event date can be used. However, unlike the standard error of a market model, implied volatilities measure the expected variability of the entire return of a stock or market index, not just the company-specific portion.<sup>11</sup> Although several experts have used the market expectation of volatility in their event studies, we propose an additional step: if we know how the stock price varies relative to the market index, we can compute the expected variability of the company-specific portion of the stock’s returns using the implied volatilities of the stock and the market index. (This approach is described in more detail in the Appendix).

The application of this method may prove problematic for the same reasons as discussed for the prior method: when the pattern of company-specific returns is alleged to be affected by the fraud even on days not directly associated with the release of news, using company-specific data from the disclosure period may be objectionable.<sup>12</sup> In the following section, we describe a method not subject to this potential problem.

### Predict volatility for disclosure dates

For situations in which there is a concern that the prolonged revelation of the fraud has contaminated the implied volatility of the firm’s stock, we propose predicting implied volatility that can be used to obtain clean event-specific standard errors. The key element to this method is that we estimate the firm-specific volatility by using the volatility of the market as a whole. This reflects the relationship documented above in connection with the heightened volatility in 2008-2009: when market volatility increases, firm-specific volatility seems also to increase. In this method, we estimate not only the market model, which relates the company’s returns to market returns, but also a model that depicts the relationship between the company’s volatility and the volatility of the market. The relationship between the volatilities in the estimation period is then used to predict the company-specific volatility in the disclosure period. The predicted company-specific implied volatility, unlike the market expectation of company-specific volatility, does not rely on the company-specific data from the event window.

There are some complexities involved in predicting company-specific volatility, as noted in academic empirical studies that analyze the volatility of broad markets as well as the idiosyncratic volatility of individual stocks.<sup>13</sup> We use a statistical approach that allows us to predict the expected return and



company-specific volatility in the event window, which can then be used in the tests of statistical significance. The procedure we employ, described in the Appendix, has the advantage of not using any data that might be contaminated by disclosures during the event period.

## Empirical illustration

To illustrate the behavior of the proposed solutions, we selected 10 companies that were constituents of the S&P 500 Index and traded from 1 July 2004- 30 June 2009.<sup>14</sup> Using the standard methodology and two of the three proposed adjustments,<sup>15</sup> we calculate the number of “statistically significant” days for four periods (1 July 2005- 30 June 2006; 1 July 2006- 30 June 2007; 1 July 2007- 30 June 2008; and 1 July 2008- 30 June 2009), using a period one year prior to the “disclosure” period as the estimation period. The average number of days found to be “statistically significant” for the three periods is summarized in Table 4.

**Table 4**

Period	Number of “Statistically Significant” Dates Using		
	Standard Methodology	Market Expectation of the Company-Specific Volatility	Index-based Prediction of the Company-Specific Volatility
1 July 2005- 30 June 2006	11.8 (4.7%)	7.7 (3.4%)	15.8 (6.3%)
1 July 2006- 30 June 2007	8.4 (3.4%)	9.1 (3.7%)	9.1 (3.6%)
1 July 2007- 30 June 2008	38.7 (15.4%)	21.6 (8.9%)	20.0 (7.9%)
1 July 2008- 30 June 2009	52.4 (20.8%)	15.8 (6.3%)	17.1 (6.8%)

While the standard methodology provides reliable estimates of the number of the significant days in the periods of relatively stable volatility, it identifies too many “statistically significant” days in the recent period of increased volatility. Using the market expectation of the company-specific volatility reduces the number of significant dates substantially, and yields approximately the expected number of statistically significant dates over the periods examined.<sup>16</sup> The final adjustment, which does not require the use of company’s implied volatility but only a volatility of a broad market index, also yields a significantly lower number of significant days compared to the use of the standard methodology during the recent volatile period.

It is worth noting that the choice and the application of the proposed adjustments for an actual case should be informed by the facts and circumstances of that case. For example, in some instances it may be prudent to use industry controls either alone or together with a broad market index.

## Summary

We argue that in times when there has been a large change in volatility, the use of the event-specific market expectation of the company-specific volatility and index-based prediction of the company-specific volatility yield more reasonable estimates of stock price inflation, and thus damages, than traditional methods. In support of this argument, we find that traditional methods of assessing statistical significance result in an implausibly high number of statistically significant days during the event period characterized by generally increased volatility.



## Appendix

### Market prediction of volatility for disclosure dates

Estimation of the expectation of company-specific volatility consists of two steps. In the first, we estimate a market model, which establishes how the company's returns vary with the returns of a market index:

$$R_t^A = \alpha + \beta R_t^M + \varepsilon_t, \quad (1)$$

where  $R_t^A$  is a return of Company A on day  $t$ ,  $R_t^M$  is a return of the broad market index on day  $t$ ,  $\alpha$  is a constant term that depicts the trend that would be observed in the company's returns if the market were flat,  $\beta$  is a coefficient capturing how the stock returns vary relative to the market index and  $\varepsilon_t$  is an error term that depicts the movement of the stock's returns that cannot be explained by the movement in a market index (which is referred to above as the company-specific portion of the stock's return).

In the second step, using  $\beta$  from the market model in (1), we calculate the expectation of the company-specific volatility using the following formula:

$$Vol_{company-specific} = \sqrt{Vol_{company-total}^2 - \beta^2 Vol_{market}^2}, \quad (2)$$

where  $Vol_{company-specific}$  is the market expectation of the company-specific volatility,  $Vol_{company-total}$  is the company's implied volatility, and  $Vol_{market}$  is the implied volatility of the market index.

### Predict volatility for disclosure dates

Suppose we are interested in assessing if a disclosure by Company A was associated with a statistically significant return. A simple approach is to first estimate a market model (1) over the estimation period. Then, over the same period, we estimate a relationship between the company's total implied volatility, and the implied volatility of the market index:

$$totVol^2(A)_t = a + bVol^2(M)_t + e_t, \quad (3)$$

where  $totVol^2(A)_t$  is the total implied volatility of Company A expressed as a variance rather than the standard deviation,  $Vol^2(M)$  is the square of the implied volatility of the market index,  $a$  is a coefficient depicting the trend in Company A's total volatility,  $b$  is a coefficient capturing how Company A's total implied volatility varies relative to the implied volatility of the market index, and  $e_t$  is an error term depicting the variation in Company A's total volatility that cannot be explained by the variation of the implied volatility of the market index.

For each date in the event window, we can calculate Company A's specific (idiosyncratic) implied volatility using the following formula:

$$predVol(A)_t = \sqrt{\hat{a} + \left(\hat{b} - \hat{\beta}^2\right) Vol^2(M)_t}, \quad (4)$$

where  $predVol(A)_t$  is the predicted company-specific (idiosyncratic) implied volatility for Company A and  $Vol(M)_t$  is the implied volatility of the market index, and  $\hat{\beta}$ ,  $\hat{a}$  and  $\hat{b}$  are estimated coefficients from the models described in (1) and (3).

Unfortunately, this simple approach does not guarantee that the predicted (total) implied volatility is positive. In equation (3), the negative relationship between the implied volatilities may cause the predicted (total) volatility for Company A to be negative. Further, the relationship between the estimated coefficients in equations (1) and (3) may be such that the predicted idiosyncratic volatility is not a real number.<sup>17</sup>

An alternative to the simple approach described above is to model the company-specific, or idiosyncratic, volatility more explicitly. In one possible specification, the error term in the market model described in (1) can be allowed to depend on the implied volatility of a broad market index.<sup>18</sup> The idiosyncratic volatility can then be estimated using a generalized autoregressive conditional heteroscedasticity model. This model may be used to predict the company-specific (idiosyncratic) volatility of Company A for the days in the event window. This method is particularly useful in instances where the implied volatility data for the company of interest is not available.



## End notes

- 
- \* The authors thank David Tabak, Brian Pastuszewski and Inez Friedman-Boyce for helping shape this paper, and Lucy Allen and Ron Miller for valuable comments and suggestions.
- <sup>1</sup> See for instance *In re Executive Telecard Ltd. Securities Litigation*, 979 F.Supp. 1021 (S.D.N.Y. 1997) and *In re Imperial Credit Industries, Inc. Securities Litigation*, 2003 WL 1563084 (C.D.Cal. 2003).
- <sup>2</sup> For simplicity's sake we will hereafter drop the alleged modifier from "alleged disclosure" and "alleged fraud" etc.
- <sup>3</sup> Event studies are a form of a classic statistical experiment: a treatment is applied to a group (the treatment group), and the outcome is compared with the outcome of the group that did not receive a treatment (the control group). If the difference between the two outcomes is found to be statistically significantly different from zero, the treatment is considered to have caused an effect. See for instance Fama, Eugene F., Jensen, Michael C., Fisher, Lawrence, and Roll, Richard W., "The Adjustment of Stock Prices to New Information," *International Economic Review*, 10(1969)1-21.
- <sup>4</sup> Market models can be used for securities other than stocks and can contain, for example, an industry index in addition to or in place of a broad market index. They can also be used to analyze a group of securities rather than just a single security.
- <sup>5</sup> In other words, the larger the standard error of the market model, the greater the abnormal return will have to be to be considered statistically significant, or different from what we would expect to see in the absence of material news.
- <sup>6</sup> The analysis examines firms in the S&P100 as of 2 March 2009. Three firms were excluded because they did not trade over the full period. Twelve of the remaining 97 firms are classified as being in the financial sector by Bloomberg's "level I classification... [which is] based on [the firm's] business or economic function and characteristics."
- <sup>7</sup> In particular, the standard error of 5.42% is roughly equal to the typical market-adjusted movement for financial companies from mid-2008 to mid-2009. The ratio of this figure to the 1.88% from the prior year equals 2.88, well above the level of 1.96 needed to show statistical significance at the 5% level.
- <sup>8</sup> In an event study, the standard test of statistical significance asks the following question: what is the probability of observing an excess return at least this large in the absence of material news? Informally, it assesses how different or unusual the return was on the event day.
- <sup>9</sup> More technically, because there is firm-related news on days in both the estimation period and the period in which we test dates for statistical significance, we would expect only 5% of the days in the testing period to be statistically significant if the proportion of news dates was unchanged (i.e., there was not a new source of material news in the event period such as one or more corrective disclosures.)

- 10 Imagine, for example, that a company announces on Monday that its previous SEC filings are no longer reliable, and is uncertain when it will file amended forms. If the amended forms are filed on Friday, it is relatively clear that Monday and Friday should be excluded from the estimation window. However, any news about the company released Tuesday through Thursday would presumably take on heightened importance to investors relative to if the company's books had been reliable. Therefore, the uncertainty regarding the company's financials may affect the volatility of its stock even on days where no news directly related to the SEC forms was issued.
- 11 Again, the company-specific portion of a stock's return is the part of the return that cannot be explained by market factors. See above, p. 2.
- 12 Suppose a company announces on Monday that it will make an important disclosure the following Friday, but does not specify if this disclosure is positive or negative. As a consequence of this announcement, the implied volatility increases, reflecting the uncertainty regarding the nature of the announced disclosure. This increase in implied volatility may lead to a finding that the disclosure was not statistically significant, even though the announcement was material.
- 13 See for instance John Y. Campbell, Lettau Martin, Malkiel Burton, Xu Yexiao, "Replication data for: Have Individual Stocks Become More Volatile? An Empirical Exploration of Idiosyncratic Risk," *The Journal of Finance* 46(2001)1-43; Spiegel, Matthew I. and Wang, Xiaotong, "Cross-sectional Variation in Stock Returns: Liquidity and Idiosyncratic Risk," *Yale ICF Working Paper No. 05-13*, 8 September 2005; Wagner, Niklas F., "Time-Varying Moments, Idiosyncratic Risk, and an Application to Hot-Issue IPO Aftermarket Returns," *Research in International Business and Finance* 18(2004)59-72; and Xu, Yexiao and Malkiel, Burton G. G., "Investigating the Behavior of Idiosyncratic Volatility," *Journal of Business* 76(2003)613-44. Chok, Jay Inghwee and Sun, Qian, "Determinants of Idiosyncratic Volatility for Biotech IPO Firms," USC Marshall School of Business Research Paper; *Financial Management*, 2007.
- 14 We use a list of companies constituting the S&P 500 Index on 2 March 2009 sorted in the alphabetical order after removing the companies that were in the S&P100 and select every 40th company. If a company did not have actively traded options or was a defendant in class action suit, that company was not included in the sample. We excluded one company based on these two criteria (Textron Inc, which is a defendant in a class action suit), and included instead the next company in the alphabetic ordering (Thermo Fisher Scientific Inc).
- 15 To properly use the first proposed adjustment and overlap the estimation and event periods, all relevant news dates need to be identified and excluded from the estimation. Due to the extensive amount of work this would entail, we did not implement that analysis for the purposes of this paper.
- 16 Calculation of market expectation of the company-specific volatility requires that the data on that company's implied volatility be available. For several companies in our sample, the implied volatility data were not available for all days during the relevant periods, which may cause a downward bias in the number of significant days (it is plausible that the day for which the implied volatility is not available is a significant day). The reported percentage of significant days is calculated based on the number of trading days for which the company implied volatility data was available.
- 17 To ensure that the predicted volatility for Company A is positive, the relationship described in equation (3) can be established between the logarithms of the two implied volatilities. In this case, the equations (3) and (4) become
- $$\ln(\text{totVol}(A)_t) = c + d \ln(\text{Vol}(M)_t) + \mu_t \quad (5)$$
- and
- $$\text{predVol}(A)_t = \sqrt{e^{(c+d \ln(\text{Vol}(M)_t))} - \hat{\beta}^2 \text{Vol}^2(M)_t} \quad (6)$$
- However, even though the log-log specification in (5) ensures that the predicted (total) implied volatility is positive, it does not ensure that the predicted idiosyncratic volatility, described in (6) is real (i.e., that the expression under the square root is positive).
- 18 In this model, the error term is no longer assumed to have a constant variance.



### About NERA

NERA Economic Consulting ([www.nera.com](http://www.nera.com)) is a global firm of experts dedicated to applying economic, finance, and quantitative principles to complex business and legal challenges. For nearly half a century, NERA's economists have been creating strategies, studies, reports, expert testimony, and policy recommendations for government authorities and the world's leading law firms and corporations. We bring academic rigor, objectivity, and real world industry experience to bear on issues arising from competition, regulation, public policy, strategy, finance, and litigation.

NERA's clients value our ability to apply and communicate state-of-the-art approaches clearly and convincingly, our commitment to deliver unbiased findings, and our reputation for quality and independence. Our clients rely on the integrity and skills of our unparalleled team of economists and other experts backed by the resources and reliability of one of the world's largest economic consultancies. With its main office in New York City, NERA serves clients from over 20 offices across North America, Europe, and Asia Pacific.

### Contacts

For more information, please contact:

**Dr. Branko Jovanovic**

Senior Consultant

+1 212 345 1972

[branko.jovanovic@nera.com](mailto:branko.jovanovic@nera.com)

*The opinions expressed herein do not necessarily represent the views of NERA Economic Consulting or any other NERA consultant. Please do not cite without explicit permission from the authors.*

# Exhibit 56



# HHS Public Access

Author manuscript

*Food Drug Law J.* Author manuscript; available in PMC 2018 October 03.

Published in final edited form as:

*Food Drug Law J.* 2017 ; 72(4): 595–635.

## When the Alpha is the Omega: *P*-Values, “Substantial Evidence,” and the 0.05 Standard at FDA

Lee Kennedy-Shaffer\*

### Abstract

A prominent feature of statistical reasoning for nearly a century, the *p*-value plays an especially vital role in the clinical testing of new drugs. Over the last fifty years, the U.S. Food and Drug Administration (FDA) has relied on *p*-values and significance testing to demonstrate the efficacy of new drugs in the premarket approval process. This article seeks to illuminate the history of this statistic and explain how the statistical significance threshold of 0.05, commonly decried as an arbitrary cutoff, is a useful tool that came to be the cornerstone of FDA decision-making.

### Introduction

The United States Food and Drug Administration (FDA) approved 16 new molecular entities between November 2016 and April 2017, according to the Drugs@FDA database.<sup>1</sup> These new drugs and biologics, whether pills, ointments, or injections, seek to treat conditions as diverse as severe genetic pediatric conditions, dermatitis, chronic kidney disease, constipation, and advanced cancers. The review files for these drugs are a maze of numbers addressing pharmacodynamics, disease incidence and prevalence, doses, results in animal models, and rates of cure or improved symptoms. But one number comes up again and again, regardless of drug class or condition treated: the *p*-value.

This number is used to distill the mountain of information in a New Drug Application (NDA) into understandable and comparable references that describe the overall quantity of evidence. Much maligned and often misinterpreted, the *p*-value plays a central role in guiding decision-making based on statistical evidence in many disciplines. Nowhere is this role more prominent than in clinical trials, where minute differences in *p*-values can mean the difference between drug approval and failure. Understanding this statistic, and the 0.05 significance level that often accompanies it, requires understanding not only its statistical meaning, but also the history of its use in statistics broadly, and clinical trials specifically. The history of FDA’s use of this statistic and this threshold value sheds light on both the

\*Lee Kennedy-Shaffer, BS, is a PhD student at the Harvard University Graduate School of Arts & Sciences, Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Ave., Building 2, 4th Floor, Boston, MA 02115. Lee\_kennedyschaffer@g.harvard.edu. The author’s involvement in the writing of this paper was partially in fulfillment of requirements for the course Food and Drug Law at Harvard Law School, taught by Peter Barton Hutt, JD. The author’s studies are supported by a grant from the U.S. National Institute of Allergy and Infectious Diseases (5T32AI007358–28). The funder had no role in the design, analysis, preparation, or decision to publish the manuscript. The opinions and analysis in the article are the author’s own.

<sup>1</sup>U.S. FOOD & DRUG ADMIN., *Drugs@FDA: FDA Approved Drug Products*, (Jul. 1, 2017) <https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm/>. The author identified sixteen “Type 1 - New Molecular Entity” approvals in this time frame from the database: Alunbrig, Austedo, Emflaza, Eucrisa, Ingrezza, Kisqali, Parsabiv, Rubraca, Rydapt, Spinraza, Symproic, Trulance, Tymlos, Xadago, Xermelo, Zejula.

outsized role they play in the contemporary drug regulatory regime and the ways in which challenges to this statistic may shape the future of FDA regulation.

## I. History of Randomized Controlled Trials IN U.S. Drug Regulation

In public health and biomedicine, the randomized, blinded, controlled trial is a paradigm of research and often the standard against which other types of evidence are measured.<sup>2</sup> FDA in particular has explicit expectations for the drug development process. Consequently, pharmaceutical and biotechnology companies are very familiar with the three phases of clinical studies, which are largely based on this paradigm.<sup>3</sup> The rise of this system and the specific rules associated with it today, however, have a long history, one that is “neither ... smooth nor ... direct” according to historian Harry Marks.<sup>4</sup> In order to understand the role of the *p*-value in the drug approval process, we begin with the source of the data, the clinical trial, and how the trial achieved its scientific and regulatory prominence.

### A. The Introduction of the Clinical Trials Paradigm

Early federal drug legislation in the United States focused on prohibiting misbranded and adulterated drugs. In 1938, the Federal Food, Drug, and Cosmetic Act (FDCA) added a prohibition on drugs that were “dangerous to health under the conditions of use prescribed in the labeling thereof.”<sup>5</sup> The law also created a premarket notification process whereby a company wishing to market a new drug submitted an application with information about the drug’s prescribed use, composition, and safety to the Secretary of Health, Education, and Welfare. Specifically, the application had to include “full reports of investigations which have been made to show whether or not such drug is safe for use.”<sup>6</sup> These investigations had to “include adequate tests by all methods reasonably applicable” to demonstrate safety.<sup>7</sup> The system did not require active approval, however; if the Secretary failed to reject the application in 60 days, it was automatically approved.<sup>8</sup> For the next 25 years, this system controlled drug approvals in the United States, without any specific reference to drug efficacy.

The mandate for “investigations” and “adequate tests” reflected a shifting paradigm in the U.S. biomedical community. While various randomized or pseudo-randomized experiments

<sup>2</sup>See, e.g., KENNETH J. ROTHMAN ET AL., MODERN EPIDEMIOLOGY § 6 (3d ed. 2008) (discussing the application of the randomized controlled trial paradigm to other types of epidemiologic evidence in order to make causal claims).

<sup>3</sup>Mark A. Goldberg et al., *Clinical Drug Evaluation and Regulatory Approval*, in PRINCIPLES OF PHARMACOLOGY: THE PATHOPHYSIOLOGIC BASIS OF DRUG THERAPY 860, 864–66 (2012).

<sup>4</sup>Harry M. Marks, The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900–1990, at 6 (1997).

<sup>5</sup>David F. Cavers, *The Food, Drug, and Cosmetic Act of 1938: Its Legislative History and Its Substantive Provisions*, 6 LAW & CONTEMP. PROBS. 2, 15 (1939). It is worth noting here that this article will focus on human drugs and biological products regulated via the submission by drug sponsors to FDA of New Drug Applications (for small-molecule drugs) and Biologics License Applications (for biologic products), as specified in Section 505 of the FDCA and regulated now by the Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, respectively, at FDA. Prior to the assignment of biologics regulation to FDA in 1972, the discussion herein applies primarily to human small-molecule drugs. In 1997, with the FDA Modernization Act, Congress explicitly mandated the harmonization of standards for NDAs and BLAs. Medical devices and animal drugs are regulated under separate frameworks and not discussed here, except for a brief discussion of statistical reviews of efficacy of medical devices, *infra* section V.C. See PETER BARTON HUTT ET AL., FOOD AND DRUG L.: CASES AND MATERIALS 135, 1124–31, 1236–38 (4th ed. 2014).

<sup>6</sup>21 U.S.C. § 355(b) (1938).

<sup>7</sup>*Id.* § 355(d).

<sup>8</sup>Cavers, *supra* note 5, at 40.

had occurred earlier, a recognizable version arose in earnest in the early twentieth century. This modern clinical trial responded to the needs of new biomedical sciences that were developing new therapies at a much faster rate than ever before. With carefully tabulated data available from hospitals and new statistical procedures ready for use, the means to scientifically test drugs became available.<sup>9</sup>

In 1915, Major Greenwood and G. Udny Yule published a paper on cholera and typhoid inoculations that specified the following three specific criteria for valid inference from clinical trials for vaccines: subjects in the inoculated and uninoculated groups must be, “in all material respects, alike”; exposure to the disease must be identical among the inoculated and uninoculated groups; and inoculation and the fact of the disease having occurred must be independent.<sup>10</sup> Other trials used similar designs throughout the 1920s and 1930s, using randomization rather than deliberate balancing in the hopes of fulfilling the first two of those criteria, and the use of placebo controls and blinding of investigators and participants—known as “double-blinding”—in the hopes of fulfilling the third.<sup>11</sup> Major independent research institutes, like the Rockefeller Institute for Medical Research, large hospitals, and medical schools took the lead in conducting these “scientific trials.”<sup>12</sup>

In the 1940s and 1950s, large-scale cooperative trials took place on penicillin and other drugs. Replicable methods and established statistical techniques became even more important as the scale of trials grew.<sup>13</sup> In 1948, epidemiologist and biostatistician A. Bradford Hill and the Medical Research Council published results from a large, multi-site study of streptomycin treatment of tuberculosis in Great Britain and, in so doing, set the standard for future efficacy trials. Noting that future investigations of therapeutic agents would “be considered valid only if based on adequately controlled clinical trials,” Hill and his colleagues detailed the methodology of their trial in some detail.<sup>14</sup> The randomization procedures and double blinding, as well as the use of multiple study sites, were particularly discussed. While results were presented in detail, few formal statistical tests were incorporated into this analysis.<sup>15</sup>

Following the publication of the Medical Research Council’s trial and a similar streptomycin experiment conducted by the U.S. Public Health Service, controlled randomized experiments became the foundation for the study of pharmaceutical safety and effectiveness. Historian Harry Marks later wrote:

Since that time, therapeutic reformers have invested controlled randomized experiments with the faith they once had in the integrity and skill of experienced researchers, in the productivity and scientific rigor of cooperative studies, and in the

<sup>9</sup>Harry F. Dowling, *The Emergence of the Cooperative Clinical Trial*, 43 TRANSACTIONS & STUD. C. PHYSICIANS PHILA. 20, 20 (1975).

<sup>10</sup>Major Greenwood, Jr. & G. Udny Yule, *The Statistics of Anti-typhoid and Anti-cholera Inoculations, and the Interpretation of Such Statistics in General*, 8 PROC. ROYAL SOC’Y MED. 113, 115–16 (1915).

<sup>11</sup>Abraham M. Lilienfeld, *The Fielding H. Garrison Lecture: Ceteris Paribus: The Evolution of the Clinical Trial*, 56 BULL. HIST. MED. 1, 14–17 (1982).

<sup>12</sup>MARKS, *supra* note 4, at 48–51.

<sup>13</sup>*Id.* at 125–26, 132–34, 138–40, 144–48.

<sup>14</sup>Streptomycin Treatment of Pulmonary Tuberculosis: A Medical Research Council Investigation, 2 BRIT. MED. J. 769, 769–71 (1948).

<sup>15</sup>*Id.* at 772–82.



ability of gate-keeping institutions such as the AMA's Council on Pharmacy and Chemistry to transform medical knowledge and practice.<sup>16</sup>

In other words, randomized clinical trials became the gold standard for evaluating drugs among those who wished to put medicine on a truly scientific basis, replacing the myriad forms of evidence clinicians and public health researchers previously considered.<sup>17</sup> Regulators soon followed the trend.

## B. The Drug Amendments of 1962 and the “Substantial Evidence” Mandate

Following the thalidomide crisis in Europe in the early 1960s and subsequent lengthy Congressional hearings on the quality of pharmaceutical studies, Congress passed the Drug Amendments of 1962.<sup>18</sup> Also known as the Kefauver-Harris Amendments, these provisions created the first mandate that new drugs be shown to be effective before approval. Specifically, a new basis for refusal of a New Drug Application (NDA) was added to FDCA section 505(d):

If the Secretary finds ... that ... (5) evaluated on the basis of the information submitted to him as part of the application and any other information before him with respect to such drug, there is a lack of substantial evidence that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the proposed labeling thereof ....<sup>19</sup>

That is, a sponsor needed to provide evidence of the drug's efficacy to gain approval. The application also now became a true premarketing affirmative approval process, rather than just an opportunity for the Secretary to reject the application. Moreover, the Secretary was empowered to begin hearings to withdraw approval if new information suggested a lack of substantial evidence of the drug's effectiveness.<sup>20</sup>

The amended section 505(d) then defines “substantial evidence” for this purpose as:

[E]vidence consisting of adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved, on the basis of which it could fairly and responsibly be concluded by such experts that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the labeling or proposed labeling thereof.<sup>21</sup>

This brief definition, appealing largely to expert opinion, makes no mention of statistical principles to be applied, nor does it direct the Secretary or FDA to promulgate any regulations outlining such principles. The only specifics it offers are that the evidence should

<sup>16</sup>Marks, *supra* note 4, at 132.

<sup>17</sup>*See id.* at 2–5 (describing the “political community” of “therapeutic reformers” and how they sought to position medicine as a scientific field).

<sup>18</sup>Robert Temple, *Development of Drug Law, Regulations, and Guidance in the United States*, in PRINCIPLES OF PHARMACOLOGY: BASIC CONCEPTS & CLINICAL APPLICATIONS 1643, 1644 (Paul L. Munson et al. eds., 1995). *See also* Jennifer Kulynych, *Will FDA Relinquish the “Gold Standard” for New Drug Approval? Redefining “Substantial Evidence” in the FDA Modernization Act of 1997*, 54 FOOD & DRUG L. J. 132–35 (1999).

<sup>19</sup>Drug Amendments of 1962, Pub. L. No. 87–781, 76 Stat. 781, 781 (codified at 21 U.S.C. § 355(d) (2016)).

<sup>20</sup>*Id.* at 784.

<sup>21</sup>*Id.* at 781.

include “adequate and well-controlled investigations” and these should include “clinical investigations,” both plural.<sup>22</sup>

Beginning immediately prior to the passage of the Drug Amendments of 1962, FDA began promulgating its own regulations on clinical trial conduct. Many of these regulations exist through the investigational new drug application (IND) procedure, which is the mechanism by which companies can legally ship experimental drugs in interstate commerce for research purposes prior to FDA marketing authorization.<sup>23</sup> Governed by Section 505(i) of the FDCA, the exemption from the standard rules of interstate drug commerce allows companies to conduct clinical trials, but it also gives FDA the power to regulate trials, a role that can be as significant as the agency desires.<sup>24</sup>

In August 1962, FDA promulgated a Notice of Proposed Rulemaking, which eventually became 21 C.F.R. Part 130, detailing the requirements of the IND process.<sup>25</sup> The final regulations were promulgated in January 1963, after passage of the Kefauver-Harris Amendments. The IND application detailed therein included a description of the three phases of trials that were expected: phase 1 on a small number of healthy subjects to determine dose, short-term toxicity, and pharmacological action; phase 2 on a limited number of patients with, or at-risk for, the target condition to determine proof of concept of efficacy; and separate, larger clinical trials in phase 3 to assess drug safety and effectiveness. Initial protocols, including investigator names and qualifications, approximate number of subjects, trial inclusion/exclusion criteria, and trial duration, were expected to be included in this application.<sup>26</sup>

The regulations did not specify strict requirements for the conduct of trials in each phase, in part to “allow flexibility in the design and execution of investigational programs.”<sup>27</sup> Meticulous record-keeping, however, was mandated. Additionally, FDA required monitoring of each trial to regularly evaluate safety and effectiveness. Specific language did concur with the statutory language requiring “investigations” (plural), by noting that phase 3 “is conducted by separate groups following the same protocol.”<sup>28</sup>

FDA’s powers thus arise not from specific language regulating trial design, but from the Commissioner’s power to revoke INDs for reasons including an unreasonable plan for

<sup>22</sup>“Substantial evidence” in other legal contexts has generally been defined as a very low standard of evidence. The Senate Report of the Kefauver-Harris Amendments further suggests that this language was used to mandate evidentiary standards that would, in a legal context, be considered fairly low. Scientifically, “substantial” can be used to denote a wide variety of levels of evidence. *See Drug Efficacy and the 1962 Drug Amendments*, 60 GEO. L. J. 185, 192–95 (1971) (detailing the legislative history of the Kefauver-Harris Amendments’ drug efficacy standard and the choice of the “substantial evidence” test); Kulynych, *supra* note 18, at 132–35, 143–47 (detailing the history of the “substantial evidence” standard up to and including the enactment of the FDA Modernization Act of 1997); Jonathan J. Darrow, *Pharmaceutical Efficacy: The Illusory Legal Standard*, 70 WASH. & LEE L. REV. 2073, 2083–88 (2013) (detailing court opinions on the “substantial evidence” standard and legal challenges to the standards employed by FDA).

<sup>23</sup>HUTT ET AL., *supra* note 5, at 674–75.

<sup>24</sup>*Id.* at 674–78.

<sup>25</sup>New Drugs for Investigational Use; exemptions from section 505(a), 27 Fed. Reg. 7990, 7990–92 (Aug. 7, 1962) (to be codified at 21 C.F.R. pt. 130.3).

<sup>26</sup>New Drugs: Procedural and Interpretative Regulations; Investigational Use, 28 Fed. Reg. 179, 180 (Dec. 31, 1962). (to be codified at 21 C.F.R. pt. 130.3); New Drugs: Investigational Drugs; Procedure Regarding Biologic Products, 5048, (to be codified at 21 C.F.R. pt. 130.3); New Drugs for Investigational Use; Foreign Shipments; Drugs Used for Diagnosing Disease, 10972–73 (to be codified at 21 C.F.R. pt. 130.3(a)).

<sup>27</sup>New Drugs: Procedural and Interpretative Regulations; Investigational Use, 179, 179 (to be codified 21 C.F.R. pt. 130.3).

<sup>28</sup>*Id.* at 180.

clinical investigations and clinical investigations not being conducted according to the submitted plan.<sup>29</sup> Through this rule, then, FDA had the power to terminate the IND exemption from the interstate commerce prohibition, thereby bringing its sponsor's ability to conduct trials to a swift end. The Committee on Public Health of the New York Academy of Medicine, amidst debate over the new regulations in 1962, noted that "it is impossible to lay down one master protocol or procedure for clinical testing" but made clear that the contemporary haphazard state of testing, full of corporate bias and cherry-picking, was not in the best interests of physicians and the public.<sup>30</sup> With these regulations, FDA had carved for itself a massive role, bearing the charge to determine for each drug what constituted "substantial evidence ... consisting of adequate and well-controlled investigations."<sup>31</sup>

## II. The Rise of the *P*-Value

The scientific trials FDA demanded, under the 1960s regulations, to support biomedical interventions required a method to summarize the accumulated data. It would be overwhelming to examine a full case report from every trial subject, so physicians and scientists running clinical trials searched for a method to summarize the data. They found such a technique in the papers of statisticians working in various fields of biology, and the *p*-value became the standard: one number to summarize the evidence from a clinical trial.

### A. The *P*-Value and Hypothesis Testing

Statistically, the *p*-value serves a very specific function. It is a measure of the compatibility of collected data with a defined scientific hypothesis. In a testing framework, a null hypothesis and an alternative hypothesis are defined. In the health sciences, the null hypothesis is often the absence of some effect, whether of a treatment or intervention, or the absence of a difference between the effects of two treatments. The alternative hypothesis is the opposite of the null hypothesis, generally that some effect or some difference is present. A statistical test is used to determine whether the evidence accords with the null hypothesis.<sup>32</sup>

Within the frequentist framework of statistical inference, there are many ways to formulate statistical tests, and they depend on both the null hypothesis and the data that will be available.<sup>33</sup> A test is generally model-dependent, meaning it relies upon some assumptions about the way in which data are generated. When data are generated according to some

<sup>29</sup>*Id.* at 182.

<sup>30</sup>Committee on Public Health, The Importance of Clinical Testing in Determining the Safety and Efficacy of Drugs, 38 BULL. N.Y. ACAD. MED. 415, 420 (1962).

<sup>31</sup>Drug Amendments of 1962, Pub. L. No. 87-781, 76 Stat. 781, 781 (1962) (codified at 21 U.S.C. § 355(d) (2016)).

<sup>32</sup>GEORGE CASELLA & ROGER L. BERGER, STATISTICAL INFERENCE 345, 345-46, 364 (1990). *See also, e.g.*, Sander Greenland et al., *Statistical Tests, P Values, Confidence Intervals, and Power: a Guide to Misinterpretations*, 31 EUR. J. EPIDEMIOLOGY 337, 338-39 (2016).

<sup>33</sup>The frequentist framework is based on hypothetical repeated sampling of data from some larger population of possible results and assessing the likelihood of the data arising under various scenarios. In contrast, the Bayesian framework considers the test statistic of interest to be a random variable and uses data and prior assumptions to determine the likelihood of various values of that parameter. *See, e.g.*, Jerzy Neyman, *Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability*, 236 PHIL. TRANSACTIONS ROYAL SOC'Y LONDON SERIES A 333, 333-47 (1937) [hereinafter Neyman, *Outline*] (describing the definition of probability used in frequentist methods for estimating parameters and testing hypotheses); Jerzy Neyman, *Frequentist Probability and Frequentist Statistics*, 36 SYNTHESIS 97, 113 (1977) (describing, decades later, the frequentist framework that had defined and been shaped by Neyman's work on hypothesis testing); ANDREW GELMAN ET AL., *BAYESIAN DATA ANALYSIS*, 3, 3-9 (2d ed. 2004) (describing the fundamentals of Bayesian inference).

probability distribution, properties of that distribution can be used to make an inference about the distribution itself. In general, hypothesis tests involve a test statistic that is a summary of the data; the mean value of some continuous outcome and the number of subjects who experienced some event are two common test statistics. A test then provides ranges of that test statistic for which the null hypothesis will be accepted or rejected.<sup>34</sup>

Within this hypothesis testing framework, the  $p$ -value, a number between 0 and 1, can be defined in several equivalent ways. The formulation most commonly used in the medical literature defines the  $p$ -value as the “probability, under the assumption of no effect or no difference, (the *null hypothesis*), of obtaining a result equal to or more extreme than what was actually observed.”<sup>35</sup> If an event is only of interest if it is more extreme in the same direction as the observed results (compared to the null hypothesis), then we use only that one-sided probability. More commonly, however, a two-sided probability is calculated that is agnostic to whether the more extreme event is in the same or opposite direction as the observed results. A (one-sided or two-sided)  $p$ -value is generally then compared to some pre-specified alpha level or significance level. If it is below the alpha level, the null hypothesis is rejected; if it is above the alpha level, the null hypothesis is not rejected. One can equivalently define the  $p$ -value, then, as the value of alpha for which the data would be on the border between rejecting and not rejecting the null hypothesis.<sup>36</sup>

The alpha level is a key part of the testing framework and has caused much controversy. Tests with a fixed alpha level can give rise to two types of errors. A Type I error occurs when the test rejects the null hypothesis despite the null hypothesis being true. A Type II error occurs when the test accepts the null hypothesis despite the null hypothesis being false. Since data are generated from a probabilistic mechanism, chance alone can lead to one of these errors. The alpha level is then the maximum probability of a Type I error; in other words, it is the maximum probability of rejecting the null hypothesis when the null hypothesis is true. The probability of not making a Type II error is called the power of a test, representing a test’s ability to detect an effect when an effect exists. A test with a small alpha level is often called a “conservative” test because it is unlikely to reject the null hypothesis when the null hypothesis is true. This often comes with a tradeoff, however; as the test is more likely to accept the null hypothesis when the null hypothesis is false, the power of the test is decreased.<sup>37</sup> For common tests, there is a maximum power that can be achieved for any given significance level.<sup>38</sup>

For determining whether a treatment has an effect on some outcome (or endpoint), tests generally rely on two main features of clinical trials: the effect size and the sample size. The effect size is some measurement of the difference in outcomes between the treatment and control arms. It could be the difference in the proportion of subjects infected by a disease after receiving a vaccine versus without vaccination, for example, or the average difference

<sup>34</sup>CASELLA & BERGER, *supra* note 32, at 359.

<sup>35</sup>Steven N. Goodman, *Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy*, 130 ANNALS INTERNAL MED. 995, 997 (1999).

<sup>36</sup>CASELLA & BERGER, *supra* note 32, at 364.

<sup>37</sup>*Id.* at 358–60.

<sup>38</sup>*Id.* at 365.

in serum cholesterol levels after taking a statin compared to taking a placebo pill. The sample size is the number of people enrolled in the trial.

In general, the larger the effect size and the more people in each arm of the trial, the smaller the  $p$ -value will be. Since the  $p$ -value is the probability of the observed effect (or a more extreme effect) occurring under the null hypothesis, a smaller  $p$ -value provides stronger evidence against the null hypothesis and in favor of the alternative hypothesis. So, the smaller the  $p$ -value is, the harder it is to explain the trial observations simply by appealing to chance variations between the outcomes among the treatment and control subjects. In this hypothesis testing framework, then, the smaller the  $p$ -value is for a specific trial, the more confident the investigator can be that the drug has an effect.

A confidence interval, which often accompanies a  $p$ -value in biomedical literature, is a range of estimates of the parameter of interest, i.e., the treatment effect. Under this same frequentist framework, a confidence interval can be calculated as the set of parameter values for the null hypothesis that, with the trial data, would result in a failure to reject the null hypothesis. That is, it is the set of parameter values under which the trial would conclude that the data support the null hypothesis. If the test is conducted with a 0.05 significance level, this is a 95 percent confidence interval. Formally, the frequentist framework does not lend itself to the statement that there is a 95 percent probability of the true effect being within this interval; rather, if the exact same experiment were conducted an infinite number of times, 95 percent of the intervals generated would include the true effect.<sup>39</sup> The connection between  $p$ -values, hypothesis test results, and confidence intervals lead to them often being used as surrogates for one another.<sup>40</sup>

## B. Early Development of the P-Value

The rise of this number to a central place of importance in a wide variety of disciplines occurred quickly, albeit with only tepid support from statisticians who pioneered its use. Mathematical statistics, including the formalization of assessing uncertainty in data accumulation, is a relatively recent scientific development. Statistics as a field in and of itself arose only in the twentieth century.<sup>41</sup> But it grew out of a long history of using probability models in games of chance, considering the uncertainty in astronomical and geological observations, and in assessing the variation in physical and social processes.<sup>42</sup>

The first known use of a statistic like the  $p$ -value to assess the likelihood of an observed effect occurring under some null hypothesis came in 1710. While he did not frame it in these terms, the Scottish physician and mathematician John Arbuthnott calculated the probability of male births exceeding female births in London for 82 years in a row. He calculated this probability under the assumption that chance governed the sex of births; that is, that each birth was independent and had equal probability of being a boy or girl. When he found this probability to be exceedingly small (about 1 in 5 septillion), he wrote: “From whence it follows, that it is Art, not Chance, that governs.”<sup>43</sup> From this rejection, it is clear that

<sup>39</sup>Neyman, *Outline*, *supra* note 33, at 347–55. *See also id.* at 403–12.

<sup>40</sup>Goodman, *supra* note 35, at 1002.

<sup>41</sup>Stephen Stigler, *The History of Statistics: The Measurement of Uncertainty before 1900* 1–4 (1986).

<sup>42</sup>*Id.* at 2–5.

Arbuthnott had decided that this miniscule probability demonstrated sex at birth was not governed by chance in an equally probable way. Following him, physicians and mathematicians studied the regularity of vital statistics (birth and death records) in many different areas with the results usually leading to rejections of chance and acceptance of a divine order.<sup>44</sup>

In 1827, French mathematician Pierre-Simon Laplace, who had already written a major treatise on probability and statistics, used a *p*-value-like statistic and a somewhat more formal hypothesis framework to analyze seasonal barometric pressure measurements. Laplace wrote that a very small value of what would today be the *p*-value “would indicate with a great likelihood that the value of *x* [the discrepancy between seasons] is not due solely to the anomalies of chance.”<sup>45</sup> Finding that very small probability (0.0000015815), Laplace concluded that “the observed discrepancy thus indicates, with an extreme likelihood, a constant cause.”<sup>46</sup> From his statements on other discrepancies that he found not significant, it appears that Laplace implicitly used a 0.01 alpha level in his hypothesis testing.<sup>47</sup>

Around the same time, another Frenchman, Siméon-Denis Poisson, extended Laplace’s methods and calculated what we would now call *p*-values and confidence intervals describing the behavior of French juries and whether they were changing due to some cause. In a manuscript in 1837, he noted that a *p*-value of 0.0897 was not strong enough “to support a belief that there has been a notable change in the causes.”<sup>48</sup> Poisson used a capital “P” to represent the probability that the observed difference in jury behavior between time periods would be less than or equal to what was observed; that is, his “P” was one minus a modern *p*-value. The notation was likely chosen simply because the value in question is a probability (fortunately, the French *probabilité* also begins with “p”).<sup>49</sup> A few pages later, Poisson notes that odds of 200 to one, or a modern *p*-value of about 0.005, are convincing enough to “believe that there was ... some real anomaly in the votes of juries.”<sup>50</sup> He goes on to make a causal claim from this probability statement, attributing the change to the French Revolution of 1830.<sup>51</sup>

Only six years later, Antoine Augustin Cournot examined differences in the proportion of male babies among various population subgroups and calculated a *p*-value, this time using “P” as the modern formulation of the quantity. It represented the *a priori* chance of the data

<sup>43</sup>John Arbuthnott, An Argument for Divine Providence, Taken from the Constant Regularity Observ’d in the Births of Both Sexes, 27 PHIL. TRANSACTIONS 186, 188–89 (1710).

<sup>44</sup>STIGLER, *supra* note 41, at 226.

<sup>45</sup>PIERRE-SIMON LAPLACE, MECANIQUE CELESTE Supp. 30 (1825 & Supp. 1827) (translating from original French: “[Q]ue la valeur de *x* n’est pas due aux seules anomalies du hasard.”).

<sup>46</sup>*Id.* at Supp. 33 (translating from original French: “L’excès observé indique donc avec une extrême vraisemblance une cause constante ....”).

<sup>47</sup>*Id.* at Supp. 35. *See also* STIGLER, *supra* note 41, at 151.

<sup>48</sup>SIMEON-DENIS POISSON, RECHERCHES SUR LA PROBABILITE DES JUGEMENTS EN MATIERE CRIMINELLE ET EN MATIERE CIVILE 373 (1837) (translating from original French: “[L]a probabilité P ... ne sont point assez considérables pour qu’on soit bien fondé à croire qu’il y ait eu quelque changement notable dans les causes.”).

<sup>49</sup>*See id.* at 372–73. *See also* STIGLER, *supra* note 41, at 189–90.

<sup>50</sup>POISSON, *supra* note 48, at 376–77 (translating from original French: “[O]n peut donc croire qu’il y a eu à cette époque quelque anomalie réelle dans les votes des jurés; et la cause de cette anomalie, qui les a rendus un peu moins sévères, a pu être la Révolution de 1830.”).

<sup>51</sup>*Id.* at 377.



attaining such a value if the chance-only process (what we now call the null hypothesis) were true.<sup>52</sup> He noted explicitly that “the importance of the deviation  $\delta$  [between two population or sample means], as given by observation, depends at once on the size of the deviation and on the size of the numbers used,” that is, the effect size and the sample size.<sup>53</sup> Cournot explicitly warned of the limits of such probabilistic statements, however, commenting on the importance of the practical meaning of effect sizes and noting that the  $p$ -value “does not at all measure the chance of truth or of error pertaining to a given judgment.”<sup>54</sup> These same concerns are still discussed over 150 years later.

By the late nineteenth century, the concept of considering the probability that observed differences in groups occurred by chance was used in psychology, economics, and other social sciences.<sup>55</sup> In 1885, Francis Ysidro Edgeworth elucidated the significance test in mathematical detail. His procedure took the differences between two populations, and divided them by a “modulus,” a function of the sample size and the spread of the individual observations.<sup>56</sup> This procedure is the same one followed today for simple hypothesis testing. Edgeworth used a very conservative test, noting that results due to chance would be “extremely improbable” if they had what we would today call a  $p$ -value of less than 0.005. The value was seen as a continuous measure of evidence, however, where various such probabilities gave indications of the strength of evidence.<sup>57</sup> Edgeworth further gave numerous examples of situations where one might use this test, ranging from population birth and death rates to economics to the flow of wasps from their nests.<sup>58</sup> This framework, testing significance using a probability model and a null hypothesis, and the  $p$ -value, even if not referred to as such, had come of age.

### C. The P-Value in the Twentieth Century: Application to Randomized Trials

Applying the concept of the probability of extreme results under the null hypothesis to biological settings and, in particular, the randomized trial, came about in the early twentieth century, due largely to the works of Karl Pearson and Ronald A. Fisher. In 1900, Pearson investigated the properties of what is now known as Pearson’s chi-squared test of independence, used to analyze tables of outcomes for different populations.<sup>59</sup> Specifically, it often tests the hypothesis of whether the probability of the outcome in population A is different from that in population B.<sup>60</sup> The test statistic, chi-squared ( $\chi^2$ ), follows a specific distribution, from which Pearson calculated probabilities.<sup>61</sup> Already seeing the utility of this

<sup>52</sup>ANTOINE AUGUSTIN COURNOT, EXPOSITION DE LA THEORIE DES CHANCES ET DES PROBABILITES 196–97 (1843). *See also* STIGLER, *supra* note 41, at 199–200.

<sup>53</sup>COURNOT, *supra* note 52, at 196 (translating from original French: “Il est évident que l’importance de l’écart  $\delta$ , comme fait d’observation, dépend à la fois de la grandeur de cet écart et de la grandeur des nombres employés ....”).

<sup>54</sup>*Id.* at 196–97 (translating from original French: “Quant à la probabilité désignée plus haut par  $\Pi$ ,... elle ne mesure point la chance de vérité ou d’erreur afférente à un jugement déterminé.”).

<sup>55</sup>STIGLER, *supra* note 41, at 260–61, 308–11.

<sup>56</sup>F. Y. Edgeworth, *On Methods of Statistics*, J. STAT. SOC’Y LONDON 181, 184–87 (1885).

<sup>57</sup>*Id.* at 185.

<sup>58</sup>*Id.* *passim*.

<sup>59</sup>Karl Pearson, On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that It Can Be Reasonably Supposed to Have Arisen from Random Sampling, 50 LONDON EDINBOROUGH & DUBLIN PHIL. MAG. J. SCI. (1900).

<sup>60</sup>CASELLA & BERGER, *supra* note 32, at 398.

<sup>61</sup>Pearson, *supra* note 59, at 175.

statistic, W. Palin Elderton produced an enlarged and improved table of  $p$ -values (again referred to simply as “P”) for  $\chi^2$  statistics with given effect sizes and sample sizes.<sup>62</sup>

Building on Pearson’s work, William Sealy Gossett—a Guinness brewery employee working on statistical methods for agricultural experimentation and quality control, who published under the pseudonym “Student”—developed an even more general method in 1908. Now known as Student’s  $t$ -distribution, the distribution arose from Gossett’s investigation of the standard deviations, a measure of the spread of data, of random samples. In his landmark 1908 paper describing this method, Gossett analyzed data from a trial of two soporific (sleep-inducing) drugs; he used examples unrelated to his brewery work to avoid revealing his identity and to avoid disclosure of any trade secrets.<sup>63</sup> He calculated  $p$ -values for the effect on hours of sleep of each drug separately, and for the performance of one drug compared to the other. In a small sample of 10 patients, Gossett found one-sided  $p$ -values for a positive effect on sleep of 0.1127 and 0.0026 for drugs 1 and 2, respectively. His standard for evidence was implicitly between these two, as he wrote that “[i]t is then very likely that 1 gives an increase of sleep, but would occasion no surprise if the results were reversed by further experiments” while describing drug 2 as almost certainly effective.<sup>64</sup> He then found a one-sided  $p$ -value of 0.0015 for the test of the hypothesis that drug 1 and drug 2 had similar effects. Of this he wrote that “odds of this kind make it almost certain that 2 is the better soporific, and in practical life such a high probability is in most matters considered as a certainty.”<sup>65</sup> It is important to note that Gossett generally reported either the probability or odds of having a smaller-than-observed effect if the treatment had no impact, so when he speaks of a high probability (“ $p$ ”) it corresponds to a low modern  $p$ -value.<sup>66</sup> Gossett did not fix cutoffs, however, and he warned against fixed levels of significance in tests three decades later, calling them “nearly valueless.”<sup>67</sup>

With Gossett’s tables of the  $t$ -distribution and Elderton’s enlarged versions of Pearson’s  $\chi^2$  tables, methods were in place to use statistics to test hypotheses in experimental trials. The synthesis of these methods and formalization of a general test statistic for hypothesis testing came from the father of modern biostatistics, Ronald Aylmer Fisher. In a 1924 paper and a 1925 monograph, Fisher used the “P” calculations of his forerunners and created a full experimental method in great generality. As is clear from the title, Fisher’s book, *Statistical Methods for Research Workers*, was explicitly directed towards practitioners of science, rather than statisticians or mathematicians, as he sought “to put into the hands of research workers, and especially of biologists, the means of applying statistical tests accurately to numerical data.”<sup>68</sup> Fisher introduces “P” early in this book, with Tables I and II dedicated to the probability of exceeding various cutoffs under the normal distribution (commonly called

<sup>62</sup>W. Palin Elderton, Tables for Testing the Goodness of Fit of Theory to Observation, 1 BIOMETRIKA (1902).

<sup>63</sup>Student (William Sealy Gossett), *The Probable Error of a Mean*, 6 BIOMETRIKA 1, 20 (1908). *See also, e.g.*, Harold Hotelling, *British Statistics and Statisticians Today*, 25 J. AM. STAT. ASS’N 186, 189 (1930) (explaining Gossett’s role at Guinness, his efforts to conceal his identity, and the role of statistics in general at the brewery).

<sup>64</sup>Student (William Sealy Gossett), *supra* note 63, at 20–21.

<sup>65</sup>*Id.* at 21.

<sup>66</sup>*Id.* at 20–21.

<sup>67</sup>*See* Stephen T. Ziliak, *The Validus Medicus and a New Gold Standard*, 376 LANCET 324, 325 (2010).

<sup>68</sup>Ronald A. Fisher, *Statistical Methods for Research Workers* 16 (1925).



a bell curve). It is here that we can trace the beginning of the alpha level of 0.05 as well. Fisher writes:

The value for which  $P = .05$ , or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a negative result only once in 22 trials, even if the statistics are the only guide available. Small effects would still escape notice if the data were insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty.  
69

This passage encapsulates many features of the modern hypothesis test. The 1.96 standard for significance of normally distributed random variables is laid out clearly, as a special case of the use of 0.05 as a reasonable cutoff point for significance. In fact, 0.05 “is convenient” for Fisher precisely because it is the  $p$ -value associated with two standard deviations from the null within a normal distribution. The normal distribution is especially important because it is a good approximation for many other distributions when sample sizes are large, a result known as the central limit theorem that was proved by Laplace in the early 19<sup>th</sup> century.<sup>70</sup> Two standard deviations also approximately corresponded to three probable errors, or “quartile distances,” of the normal distribution, a measure commonly used during, and prior to, Fisher’s era but that eventually was fully replaced by the standard deviation.<sup>71</sup>

It is important to note that Fisher uses a lower value of “P” to indicate more evidence against the null hypothesis, which is the modern formulation of the  $p$ -value that is most common. Fisher’s 1.96 standard is based on the two-sided tests he prefers, although he notes that “P” can be divided by two for a one-sided test.<sup>72</sup> He also clearly states throughout his works that, to make more precise inference, the investigator should collect more data rather than lowering significance thresholds.<sup>73</sup>

Fisher is not entirely precise with his wording interpreting the alpha level, however, especially with regard to the 1 in 22 trials. In fact, of all studies *for which there is no true effect*, an average of 1 in 20 (or 22, depending on the exact cutoff used) will display a significant effect at this threshold. But that is not the same as saying that of all trials we examine, no more than 1 in 20 will be false indications.<sup>74</sup> To obtain this latter probability, one would need to know the proportion of all trials examined for which a true effect existed.

<sup>69</sup>*Id.* at 47.

<sup>70</sup>STIGLER, *supra* note 41, at 136–38. The convenience of these standard deviations for the normal were in fact noticed earlier for the special case of the binomial distribution, where each outcome is either a success or failure, and the total result is a count of successes. Abraham De Moivre published the probabilities for being within one, two, and three standard deviations of the mean of a normal distribution in 1733 (translated by the author into English in 1738). His values are very accurate, giving the now-familiar 0.954 for the probability of being within two standard deviations. ABRAHAM DE MOIVRE, THE DOCTRINE OF CHANCES; OR, A METHOD OF CALCULATING THE PROBABILITIES OF EVENTS IN PLAY 238–40 (2d ed. 1738). *See also* STIGLER, *supra* note 41, at 82.

<sup>71</sup>FISHER, *supra* note 68, at 47–48.

<sup>72</sup>*Id.* at 47.

<sup>73</sup>*Id.*

<sup>74</sup>*Id.*

Fisher proceeds to apply his method to other distributions of data and other summary measures that are being tested. For example, he re-defines the value in a chapter on the chi-squared distribution: “P ... is therefore the probability that  $\chi^2$  shall exceed any specified value.”<sup>75</sup> Fisher reiterates his 0.05 cutoffs here:

If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05, and consider that higher values of  $\chi^2$  indicate a real discrepancy.<sup>76</sup>

Fisher uses this formulation, both the 0.05 “significance” threshold and a threshold around 0.01 or 0.02 as “strong evidence,” throughout the text in his data analysis examples.

*Statistical Methods* proved to be groundbreaking both for its harmonization of different distributional methods into the *p*-value significance testing framework and its presentation of new methodology that enabled testing to be performed in a wider range of settings. Fisher introduced many forms of exact tests, which are more accurate in small-sample cases, as well as what is now called the *F*-distribution for the analysis of variances, all using *p*-values.<sup>77</sup> He also discussed randomization principles required for significance tests to be valid for experimental results, work upon which he would later expand.<sup>78</sup>

Fisher’s contributions did not end with *Statistical Methods*; in 1935, he wrote a monograph entitled *The Design of Experiments*. In the introduction, Fisher describes the importance of both satisfactory and standardized statistical procedures and elucidates his principles of experimentation.<sup>79</sup> Randomization plays the key role throughout his text, but adjustment for confounding factors (other variables that affect both the probability of receiving a certain treatment and the probability of having the outcome in question), appropriate sample sizes, and determining the mechanisms of chance are treated in detail as well. Notably, Fisher again suggests a 0.05 significance standard (stating that it is “usual and convenient for experimenters to take 5 per cent. as a standard level of significance”), but allows that investigators may wish to specify their own standards based on their purpose and how “exacting” they wish to be.<sup>80</sup>

With a general statistical theory for significance testing and principles for experimental design laid out, Fisher completed his trio of applied biostatistics monographs with *Statistical Tables for Biological, Agricultural, and Medical Research*, written with Frank Yates and published in 1938. In it, the two statisticians presented thirty-four tables of calculated results from common distributions and tests laid out in Fisher’s prior works.<sup>81</sup> The book also presented numerous examples of the use of the tables. Most of these focus on the calculation of “P” or construction of a confidence interval from an experiment, demonstrating Fisher’s belief in the importance of these measures for a wide variety of statistical work.<sup>82</sup> Equally

<sup>75</sup>*Id.* at 78.

<sup>76</sup>*Id.* at 79.

<sup>77</sup>*Id.* at 176–210.

<sup>78</sup>*Id.* at 224–29.

<sup>79</sup>Ronald A. Fisher, *The Design of Experiments* 2–3 (1935).

<sup>80</sup>*Id.* at 15–16.

<sup>81</sup>Ronald A. Fisher & Frank Yates, *Statistical Tables for Biological, Agricultural, and Medical Research* 25–90 (1938).

important, these tables, as they had only limited space for values, almost all exclusively gave values that would be required for determining the 0.05 and 0.01 levels of significance. For the tests of significance for two-by-two contingency tables—tables commonly employed in drug trials showing the number of subjects with each of two outcomes in each of two treatment arms—only these two levels are given.<sup>83</sup> Future medical statisticians wrote of the importance this had in cementing the status of these values.<sup>84</sup>

How common these standards were in Fisher's time, however, is not entirely clear. Fisher certainly preferred them, and his student L.H.C. Tippett also invoked the 0.05 standard throughout his influential 1931 book *The Methods of Statistics*, calling "the 0.05 level ... a good compromise" between the two types of error.<sup>85</sup> Tippett also, as is now common, specifically referred to *p*-values less than 0.05 as "*statistically significant*," although he immediately pointed out that his "choice of 0.05 is quite arbitrary" but "in common use."<sup>86</sup> However, Fisher's "statement that it is *usual* for research workers to adopt a 5 per cent significance level in the same context," wrote Lancelot Hogben, "is true only of those who rely on the many rule of thumb manuals expounding Fisher's own test prescriptions."<sup>87</sup> Whether due to a philosophical decision or simply the convenience of Fisher's tables, the level did become a common benchmark over the succeeding decades, referred to even by Fisher's antagonists, Jerzy Neyman and Egon Pearson, and subsequently incorporated as the "most sacred" threshold in papers and textbooks in biology and the social sciences.<sup>88</sup>

With these three books, Fisher crafted a robust framework for significance testing. Randomized experiments of many forms, testing almost any specific parameter, could be translated into test statistics with known distributions. With data from almost any common experiment, an investigator could look up the relevant table and calculate a *p*-value or a confidence interval for the hypothesis or parameter of interest. The stage was thus set for the rise of randomized experiments in the 1940s, as described *supra* section II.A. Perhaps Fisher's greatest contribution was his work to make statistical methods available to investigators without statistical training. But that very work also contributed to the misuse of statistics and overreliance on *p*-values that would later lead to crises of confidence in those methods.

#### D. Neyman, Pearson, and the Formal Hypothesis Testing Framework

As Fisher was elucidating this view of *p*-values as a continuous measure of evidence, other statisticians were constructing a more formal version of hypothesis testing, one where the results could lead only to a decision to reject or accept a hypothesis. Jerzy Neyman and Egon Pearson, son of Karl Pearson, published their key paper in 1932, in which they described this framework, later known as the Neyman-Pearson method. In their framework, investigators explicitly specify both a null hypothesis and alternative hypothesis, and in the

<sup>82</sup>*Id.* at 1–22.

<sup>83</sup>*Id.* at 37.

<sup>84</sup>*See, e.g.*, Donald Mainland, *Statistical Ward Rounds*—2, 8 CLINICAL PHARMACOLOGY & THERAPEUTICS (1967).

<sup>85</sup>L. H. C. Tippett, *The Methods of Statistics: An Introduction Mainly for Workers in the Biological Sciences* 51 (1931).

<sup>86</sup>*Id.* at 48.

<sup>87</sup>Lancelot Hogben, *Statistical Theory* 495 (1957).

<sup>88</sup>James K. Skipper, Jr. et al., *The Sacredness of .05: A Note Concerning the Uses of Statistical Levels of Significance in Social Science*, 2 AM. SOC. 16, 16 (1967). *See also* Mainland, *supra* note 84.

end reject or accept the null hypothesis.<sup>89</sup> This decision results in one of three outcomes: a correct determination, a Type I error (incorrectly rejecting the null), or a Type II error (incorrectly accepting the null). The error that is worse “will depend upon the consequences of the error.”<sup>90</sup> Critically, the authors recognize that no single decision can be classified as incorrect or correct from purely statistical data, but rather their procedure describes “rules to govern our behavior . . . , in following which we insure that, in the long run of experience, we shall not be too often wrong.”<sup>91</sup>

For any data or test statistic arising from data, it is impossible to minimize both types of error. Neyman and Pearson prove, however, that if there is a specified maximum level allowed for probability of a Type I error (denoted epsilon in the paper, but now commonly denoted alpha), then there is a test that minimizes the Type II error for every true parameter.<sup>92</sup> This test is now called the uniformly most powerful test, with power denoting the probability of rejecting the null hypothesis when it is indeed false. The alpha level will depend on the investigator’s judgment in weighing the consequences of Type I and Type II errors, but Neyman and Pearson present the familiar 0.05 and 0.01 levels as examples in their article.<sup>93</sup> Neyman and Pearson go on to show that for many common distributions, the test corresponds to “the ordinary test for the significance of a variation in the mean of a sample”<sup>94</sup>; that is, it can be constructed by determining the *p*-value in the way done by Karl Pearson, William Sealy Gossett, and R.A. Fisher. The difference from Fisher’s approach is that the *p*-value is no longer a continuous measure of evidence, but rather a test statistic to be compared to a strict cutoff from which a decision is made on the hypotheses.

While Pearson later acknowledged the debt to Fisher’s tables and to his stipulation of 0.05 and 0.01 significance levels, a fierce debate raged between Fisher and his two contemporaries about the relative benefits of their frameworks.<sup>95</sup> Fisher’s main objections arose from Neyman and Pearson’s rejection of the *p*-value as a continuous measure and their emphasis on power. Neyman and Pearson essentially claim that a test with enough power provides evidence *for* the alternative hypothesis rather than simply *against* the null hypothesis. If one of the two must be true, this is a fair statement, but Fisher warned of the many ways in which a null hypothesis can fail, and so did not want to place as much emphasis on the pre-specified alternative.<sup>96</sup> But the Neyman-Pearson framework had a valuable role in decision-making. Fisher himself noted the distinction, describing his methods as a tool for accumulating knowledge rather than for making a final decision.<sup>97</sup>

<sup>89</sup>Jerzy Neyman & Egon S. Pearson, *On the Problem of the Most Efficient Tests of Statistical Hypotheses*, 231 PHIL. TRANSACTIONS ROYAL SOC’Y LONDON SERIES A CONTAINING PAPERS OF A MATHEMATICAL OR PHYSICAL CHARACTER, 289, 294–95 (1933).

<sup>90</sup>*Id.* at 296.

<sup>91</sup>*Id.* at 291.

<sup>92</sup>*Id.* at 302.

<sup>93</sup>*Id.* at 303, 305.

<sup>94</sup>*Id.* at 304.

<sup>95</sup>See Erich L. Lehmann, *The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?*, 88 J. AM. STAT. ASS’N 1242, 1242–44 (1993); Johannes Lenhard, *Models and Statistical Inference: The Controversy between Fisher and Neyman-Pearson*, 57 BRIT. J. PHIL. SCI. 69, 70, 81 (2006).

<sup>96</sup>Lehmann, *supra* note 95, at 1244–45.

<sup>97</sup>Ronald A. Fisher, *Statistical Methods and Scientific Inference* 99 (1956).

## E. The Biomedical Synthesis

In biomedical research, these two approaches—the Fisher “weight of evidence”  $p$ -value and the Neyman-Pearson formal hypothesis test—have often been combined within a larger frequentist framework. An investigator will specify their null and alternative hypotheses, as well as a pre-specified alpha level (generally, 0.05). She will then calculate a  $p$ -value from the data and an assumed statistical model. This  $p$ -value will be compared to the alpha to determine “significance” and the null will be accepted or rejected. The  $p$ -value will also be presented as a continuous measure, often termed “statistically significant” if it is under 0.05 or “highly statistically significant” if it is under 0.01; some reference to the degree of significance may also be made by comparing the  $p$ -value to various levels.<sup>98</sup> The rise of other statistical models (for example, the Cox proportional hazards model for survival time) that made use of these frameworks and allowed for calculations of  $p$ -values and confidence intervals,<sup>99</sup> and the rise of computer software that made calculations of  $p$ -values easier and more exact than using tables, allowed the  $p$ -value and the hypothesis or significance testing framework to take precedence in biomedical research.

This synthesis can be seen to some degree in the works of Fisher and of Neyman and Pearson, as well as in the works of statisticians who came soon after them. W. Edwards Deming, in his 1943 book *Statistical Adjustment of Data*, calculated  $p$ -values—and, in fact, appears to be the first to use the term “ $P$  value”<sup>100</sup>—and suggested using  $p$ -values from repeated experiments as measures of the quantum of evidence against the null hypothesis. He also recommended the use of “statistical significance” as an inferential method. But, he warned, “[s]tatistical ‘significance’ by itself is not a rational basis for action.”<sup>101</sup>

Indeed, by the early 1950s, statistics held a prominent place in clinical trials. In the landmark 1948 study of streptomycin by the Medical Research Council, discussed *supra* section II.A, both chi-squared tests and  $t$ -tests were used to evaluate the responses to the drug and compare the control and treated groups.<sup>102</sup> At the end of the article, the authors confidently state that “[t]he difference in mortality between the two groups is statistically significant.”<sup>103</sup> Interestingly, the authors do not report the calculated  $p$ -value for any test.

Leading clinical journals soon began to note the importance of such statistical arguments as well. A 1950 editorial in the *Journal of the American Medical Association* (JAMA) posed the question “Are Statistics Necessary?” The answer was an unqualified yes: “If [an investigator] is developing a new therapy, he must know how to set up fourfold tables

<sup>98</sup>Shein-Chung Chow & Jen-Pei Liu, *Design and Analysis of Clinical Trials: Concepts and Methodologies* 73–75 (2d ed. 2004). *See also* Goodman, *supra* note 35, at 1000–01; Stuart J. Pocock, *Clinical Trials: A Practical Approach* 204–206 (1983).

<sup>99</sup>*See, e.g.*, David R. Cox, *Regression Models and Life-Tables*, 34 J. ROYAL STAT. SOC’Y SERIES B (METHODOLOGICAL) 187, 187–89 (1972).

<sup>100</sup>W. Edwards Deming, *Statistical Adjustment of Data* 30 (1943). *See also* H. A. David & A. W. F. Edwards, *Annotated Readings in the History of Statistics* 223 (2001).

<sup>101</sup>DEMING, *supra* note 100, at 30. This argument was common during the development of these methods and remains so to this day. In its simplest and most common form, it warns of statistical significance supplanting clinical significance or some assessment of effect size altogether as a standard. This is discussed *infra* section V. *See also, e.g.*, FISHER, *supra* note 79, at 194 (ascribing to the experimenter the duty of determining what “observational discrepancy ... interests him”); STEPHEN T. ZILIAK & DEIRDRE N. MCCLOSKEY, *THE CULT OF STATISTICAL SIGNIFICANCE: HOW THE STANDARD ERROR COSTS US JOBS, JUSTICE, AND LIVES* 33–42 (2011) (warning of the “sizeless” reliance of statistical inference in several scientific disciplines).

<sup>102</sup>*See* D. D. Reid, *Statistics in Clinical Research*, 52 ANNALS N.Y. ACAD. SCI. 931, 933 (1950).

<sup>103</sup>*Streptomycin Treatment of Pulmonary Tuberculosis: A Medical Research Council Investigation*, *supra* note 15, at 782.

comparing treated with untreated subjects and must know how to compute the probability that apparently favorable results were accidental.”<sup>104</sup> In this language of probability of results due to chance, we see the familiar conceptualization of the  $p$ -value arise once again. An article in the *Annals of the New York Academy of Sciences* similarly called for quantification of clinical trial results and noted that “[s]tatistical reasoning is needed as soon as that experiment is conceived.”<sup>105</sup> Additionally, an article in *JAMA* on the use of controls in medical research presupposed that statistical tests would form the basis of evidence of therapeutic effectiveness, urging clinicians to use randomization and untreated controls as “the basis for statistical comparison” and significance testing.<sup>106</sup>

In the context of drug approvals, both testing frameworks offer advantages. In order to approve a drug, FDA must decide whether the trials provide “substantial evidence that the drug will have the effect it purports or is represented to have,”<sup>107</sup> so a decision-making framework with a strict cutoff is desired. The ability to calculate power is useful to drug sponsors, who have to decide how many patients to enroll in a trial, i.e., how big the sample size will be, in order to demonstrate a true effect exists. But since FDA specifically noted that trial design and evaluation require case-by-case methods, a continuous measure such as Fisher’s  $p$ -value gives a valuable tool to assess the quantity of evidence that a drug has an effect. This tension between the goals of finding as many true effects as possible while not ascribing truth to too many false effects has persisted from its roots in the statistical literature of the 1930s. Today, it survives as the tension that has characterized FDA’s drug approval process since 1962, i.e., how to approve all beneficial drugs without approving ineffective drugs.

### III. FDA Guidance on Statistical Considerations in Clinical Trials

By 1962, when Congress passed the Kefauver-Harris Amendments, statistical methodologies, including hypothesis testing via the  $p$ -value, had been combined with the principles of sound experimental design to create an overall structure for clinical drug testing in humans. These principles were not common, much less ubiquitous, in the drug development process, however. Robert Temple, Director of the Office of Medical Policy at the FDA Center for Drug Evaluation and Research, noted that studies submitted in the 1960s often had “no protocol at all. There was almost never a statistical plan.”<sup>108</sup> Dr. Louis Lasagna, a prominent pharmacologist at Johns Hopkins University, made a similar point in Senate testimony in 1959, testimony that proved an important precursor to that for the Kefauver-Harris Amendments: “Adequately controlled comparisons of these drugs are almost impossible to find.”<sup>109</sup> While he referred specifically to corticosteroids, Dr. Lasagna made clear that his comments generally held true for the drug industry as a whole.

<sup>104</sup>Editorial, *Are Statistics Necessary?*, 143 J. AM. MED. ASS’N 1260, 1260 (1950).

<sup>105</sup>Reid, *supra* note 102, at 931.

<sup>106</sup>Otho B. Ross, Jr., *Use of Controls in Medical Research*, 145 J. AM. MED. ASS’N 72, 72 (1951).

<sup>107</sup>Pub. L. No. 87–781, 76 Stat. 781 (1962) (codified at 21 U.S.C. § 355(d)).

<sup>108</sup>Robert Temple, *How FDA Currently Makes Decisions on Clinical Studies*, 2 CLINICAL TRIALS 276, 276 (2005).

<sup>109</sup>*Hearings Before the Subcomm. on Antitrust and Monopoly of the Senate Comm. on the Judiciary (pursuant to S. Res 57)*, 86th Cong. 8138–39 (1959) (testimony of Louis Lasagna, M.D., Johns Hopkins University School of Medicine).



In 1969, FDA sought to overcome this lack of formal scientific and regulatory rigor. The Administration thus began in earnest its role in standardizing clinical trial design protocols. The agency promulgated regulations requiring specific elements in a protocol submitted for an IND. Among these was a “summary of statistical methods used in analysis of the data derived from the subjects.”<sup>110</sup> Soon after this, analysis plans became more common and more scientific. Temple noted that all sponsors “came to believe that trials should have a prospectively defined and identified endpoint, a real hypothesis and an actual analytical plan.”<sup>111</sup> As a result, FDA began using these statistical tests in decision-making, and the 0.05 standard became enshrined in U.S. drug development.<sup>112</sup>

#### A. The Rise of the 0.05 Standard in Biomedicine

The usual FDA paradigm traces its roots directly to the regulations implementing the Kefauver-Harris Amendments. The plural “adequate and well-controlled investigations” and subsequent guidance established a standard that two trials following the same protocol should generally be used.<sup>113</sup> Regulations promulgated in 1970 demanded that studies provide “a comparison of the results of treatment or diagnosis with a control in such a fashion as to permit quantitative evaluation.”<sup>114</sup> These regulations were somewhat delayed due to prolonged lawsuits between drug manufacturers and FDA over the content of the regulations. Most of these focused on legal principles and had little to do with the substantive definitions in the regulations, but the definitions put forth for “adequate and well-controlled investigations” were considered. In *Pharmaceutical Manufacturers Association v. Richardson*, the U.S. District Court for the District of Delaware found that the requirements, including the “quantitative evaluation” rule and the use of appropriate methods of data analysis, were “minimal requirements for any valid objective study” and thus “not arbitrarily rigid.”<sup>115</sup> The regulations, noted the court, “describe broad scientific standards” and still retain flexibility for the sponsor and investigator.<sup>116</sup> The regulations were thus upheld.

The regulations, once finally in force, clearly indicated the use of statistical techniques to account for the possibility of random deviations in the presence of no treatment effect. Following common clinical trial practice at the time, this involved the use of significance testing with two-sided tests at the customary significance (or alpha) level of 0.05.<sup>117</sup> While this level for controlling Type I error was not specified in regulations, it was discussed in the biomedical and statistical literature and came to be understood as the customary level, with any deviations from that needing to be pre-specified and defended in the analysis plan.<sup>118</sup> Lancelot Hogben wrote in 1957 that “[c]ontemporary literature of therapeutic and

<sup>110</sup>34 Fed. Reg. 14596, 14597 (Sept. 10, 1969).

<sup>111</sup>Temple, *supra* note 108, at 276.

<sup>112</sup>See DANIEL P. CARPENTER, REPUTATION AND POWER: ORGANIZATIONAL IMAGE AND PHARMACEUTICAL REGULATION AT THE FDA 269–97 (2010) (detailing FDA’s methods of standardizing clinical trial regimes).

<sup>113</sup>Pub. L. No. 87–781, 76 Stat. 781 (1962) (codified at 21 U.S.C. § 355(d)). See also 28 Fed. Reg. 179, 180 (Dec. 31, 1962).

<sup>114</sup>35 Fed. Reg. 7250, 7251 (Apr. 30, 1970).

<sup>115</sup>Pharmaceutical Mfr. Ass’n v. Richardson, 318 F. Supp. 310–11 (D. Del. 1970).

<sup>116</sup>*Id.* at 311. The Court did avoid further defining any of the evidentiary requirements, and refrained from weighing in on the questions of the quantum of evidence and number of required trials. See *Drug Efficacy and the 1962 Drug Amendments*, *supra* note 22.

<sup>117</sup>Bruce A. Barron & Samuel C. Bukantz, *The Evaluation of New Drugs: Current Food and Drug Administration Regulations and Statistical Aspects of Clinical Trials*, 119 ARCHIVES INTERNAL MED. 547, 553 (1967).

prophylactic trials is an uninterrupted record of Chi Square tests for  $2 \times 2$  tables to test the null hypothesis that there is no treatment difference,” referring to one of the main Fisherian methods expounded in the statistician’s works.<sup>119</sup> In using these methods, “the overwhelming majority of research workers in the biological field ... rely largely on rule of thumb procedures set forth in a succession of manuals modeled on *Statistical Methods for Research Workers* by R. A. Fisher,” including the 0.05 significance level.<sup>120</sup> Indeed, testing at the 0.05 alpha level became so commonplace in clinical trials that reporting of actual *p*-values was frequently replaced by reporting of only the result of the test, to the chagrin of some biostatisticians.<sup>121</sup>

Reviews of the major drug trials of the time also showed that 0.05 had become accepted practice. In a systematic review of 146 antidepressant drug studies conducted between 1958 and 1972, Jeffrey Morris and Aaron Beck used reported results indicating significant improvement against placebo at the 0.05 significance level.<sup>122</sup> In an analysis of the very large University Group Diabetes Program trial, 0.05 became the standard for significance of a wide variety of outcomes, including the primary outcomes of fatal and nonfatal vascular complications.<sup>123</sup> A review of original research articles in the *New England Journal of Medicine* in 1978 and 1979 found that nearly three-fourths of them used more than descriptive statistics, with *p*-values from hypothesis tests among the most frequent statistical techniques.<sup>124</sup> Major textbooks also described the use of *p*-values and significance tests, forming a key part of the education of new clinical investigators.<sup>125</sup>

The appeal by FDA to common practice among medical statisticians is not surprising, especially given the statute’s own appeal to “experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved.”<sup>126</sup> The practicing trialists, physicians, and statisticians were presumably the experts to whom this statute referred. Significance tests and the 0.05 standard had long since made the leap from the statistical works of Fisher, Neyman, and Pearson, to the medical literature. Dr. Donald Mainland, professor of medical statistics at New York University Medical Center, wrote of the significance level in the *Journal of Clinical Pharmacology and Therapeutics* in 1963.<sup>127</sup> A few years later, he began a series of commentaries in the same journal known as “Statistical

<sup>118</sup> *Id.* at 553. See also Jerome Cornfield, *Sequential Trials, Sequential Analysis, and the Likelihood Principle*, 20 AM. STAT. 18, 18 (1966); Robert T. O’Neill, *P-Values, Hypothesis Testing, and Reproducibility: An FDA Perspective*, HARV. UNIV. (Apr. 2, 2017), [https://catalyst.harvard.edu/pdf/biostatseminar/O’Neill\\_Slides.pdf](https://catalyst.harvard.edu/pdf/biostatseminar/O’Neill_Slides.pdf) [<https://perma.cc/P7WK-U6G6>].

<sup>119</sup> Hogben, *supra* note 87, at 497.

<sup>120</sup> *Id.* at 487. Egon Pearson also generally included the 0.05 and 0.01 levels in the tables he published, thus making them generally accessible for the variety of tests investigators wished to conduct in the precomputer era. See, e.g., Egon S. Pearson, *A Further Development of Tests for Normality*, 22 BIOMETRIKA 239, 240 (1930).

<sup>121</sup> Stuart J. Pocock et al., *Statistical Problems in the Reporting of Clinical Trials: A Survey of Three Medical Journals*, 317 NEW ENGL. J. MED. 426, 431 (1987).

<sup>122</sup> Jeffrey B. Morris & Aaron T. Beck, *The Efficacy of Antidepressant Drugs: A Review of Research (1958 to 1972)*, 30 ARCHIVES GEN. PSYCHOL. 667, 667 (1974).

<sup>123</sup> Alvan R. Feinstein, *Clinical Biostatistics VIII. An Analytic Appraisal of the University Group Diabetes Program (UGDP) Study*, 12 CLINICAL PHARMACOLOGY & THERAPEUTICS (1971).

<sup>124</sup> John D. Emerson & Graham A. Colditz, *Use of Statistical Analysis in The New England Journal of Medicine*, 309 NEW ENGL. J. MED. (1983).

<sup>125</sup> See, e.g., Mainland, *supra* note 84, at 348; POCOCK, *supra* note 98, at 197–206; James H. Ware et al., *P Values, in MEDICAL USES OF STATISTICS* 181, 181 (John C. Bailar III & Frederick Mosteller eds., 2d ed. 1992).

<sup>126</sup> Pub. L. No. 87–781, 76 Stat. 781 (1962) (codified at 21 U.S.C. § 355(d)).

<sup>127</sup> Donald Mainland, *Commentary: The Significance of “Nonsignificance”*, 4 CLINICAL PHARMACOLOGY & THERAPEUTICS (1963).



Ward Rounds,” which addressed the statistical questions of clinical trialists and physicians.<sup>128</sup> His first substantive commentary directly addressed significance testing in drug trials for efficacy, though not always favorably, and was replete with the 0.05 standard.<sup>129</sup> In a similar series entitled “Clinical Biostatistics,” Dr. Alvan Feinstein of the Yale School of Medicine continued to use 0.05 as the threshold for significance, even while proposing newer methods of trial design and statistical analysis.<sup>130</sup> As such a key part of the education, work, and publications of trialists, it is natural that *p*-values came to be the accepted form of evidence for the experts making decisions at FDA.

## B. Initial FDA Implementation of Statistical Standards

One of the first opportunities for FDA to implement this standard was in the Drug Effectiveness Study, which began in the mid-1960s with the task of determining the effectiveness of drugs on the market prior to the enactment of the Kefauver-Harris Amendments. The National Academy of Sciences/National Research Council undertook the review of thousands of NDAs, classifying each on a six-category scale based on the evidence for the drug’s effectiveness.<sup>131</sup> The requirements for evidence were meant to be the same as they would be for new drug applications going forward.<sup>132</sup> Nonetheless, in part because of, as one reviewer put it, a general lack of “statistically valid experimental evidence,” this task became quite difficult.<sup>133</sup> The final report of the Drug Efficacy Study is thus more useful in providing intuition on how evidence would be viewed under the new regime rather than the specific standards employed.

In comments on the process, Drug Efficacy Study reviewers noted many issues with previously conducted clinical trials, encouraged FDA to set forth clear standards, and requested that the agency work with sponsors going forward to ensure appropriate design and analysis.<sup>134</sup> One reviewer specifically brought up statistical principles of design, noting the need for trials to be designed “so that the level of significance of differences between efficacy and spontaneous regression” of disease can be determined.<sup>135</sup> The reviews, publicly available, “identified hundreds, perhaps thousands, of examples of inappropriate, after-the-fact data subsetting ..., and essentially every other design and statistical ‘crime’ that could be committed.”<sup>136</sup> In the reviews themselves, clear standards were not always set, but the Panel on Antiemetic Drugs (drugs that combat nausea) did explicitly call for statistical analyses “to determine whether observed differences between test and control groups are likely to be caused merely by chance.”<sup>137</sup> Given trial practice at the time, this meant significance testing with a pre-specified alpha level. The general principles and issues

<sup>128</sup>Donald Mainland, *Statistical Ward Rounds—I*, 8 CLINICAL PHARMACOLOGY & THERAPEUTICS 139, 139 (1967).

<sup>129</sup>Mainland, *supra* note 84, at 349–51.

<sup>130</sup>Alvan R. Feinstein, *Clinical Biostatistics V. The Architecture of Clinical Research (concluded)*, 11 CLINICAL PHARMACOLOGY & THERAPEUTICS 755, 759 (1970).

<sup>131</sup>See HUTT, *supra* note 5, at 776–77.

<sup>132</sup>35 Fed. Reg. 7250, 7250–51 (Apr. 30, 1970).

<sup>133</sup>NAT’L RESEARCH COUNCIL, DRUG EFFICACY STUDY: FINAL REPORT TO THE COMMISSIONER OF THE FOOD AND DRUG ADMINISTRATION 61–62 (1969).

<sup>134</sup>*Id.* at 92–96.

<sup>135</sup>*Id.* at 96.

<sup>136</sup>Temple, *supra* note 18, at 1656.

<sup>137</sup>NAT’L RESEARCH COUNCIL, *supra* note 133, at 116.

identified in the Drug Efficacy Study would go on to serve as a basis for FDA guidance in post-1962 NDA considerations as well.

In the late 1970s, statistical tests were showing up in other legal areas as well. In the grand jury discrimination case *Castaneda v. Partida*, the U.S. Supreme Court conducted a hypothesis test and referred to “a general rule” that if the observed data yield a statistic “greater than two or three standard deviations” from the expectation of the null hypothesis, the hypothesis “would be suspect to a social scientist.”<sup>138</sup> These standards are equivalent to approximately a 0.05 or 0.01 alpha level for significance testing. The Court supported its use of what are essentially *p*-values in discrimination cases with reference to a 1966 article in the *Harvard Law Review*.<sup>139</sup> In it, Michael Finkelstein calculates *p*-values for several jury discrimination cases and finds them less than 0.05, “the value most commonly used by statisticians.”<sup>140</sup> While the statistical test alone does not provide proof of discrimination, writes Finkelstein, it does demonstrate that the proportion of black jury members was not consistent with the racial makeup of the area.<sup>141</sup> These criteria were applied again in the teacher employment discrimination case *Hazelwood School District v. United States* to find statistically significant evidence of discrimination. In this case, however, the Court stated that “these observations are not intended to suggest that precise calculations of statistical significance are necessary in employing statistical proof.”<sup>142</sup> While the evidentiary standards in criminal and civil cases are different than those employed by FDA, these cases nonetheless suggest that statistical testing and the 0.05 and 0.01 significance levels had, to some degree, been endorsed by the highest court in the land.<sup>143</sup>

Beginning in the 1970s, FDA expanded its biostatistical corps, which comprised at the time “a few statisticians” with a “modest at best” role in drug review.<sup>144</sup> Statisticians came to contribute “at all levels of review, not only to the review of clinical data and study design, but to the review of” various early-phase studies.<sup>145</sup> This was an ongoing process, however, as Louis Lasagna, who played a major role in formalizing clinical trials and guiding federal drug policy, noted in 1989. He said that even then, a quarter of a century after the passage of the Kefauver-Harris Amendments, FDA was still expanding its role in reviewing clinical trial protocols in the IND submission process.<sup>146</sup>

A high-profile role for the biostatisticians came in 1980, when FDA reviewed a new claim by Ciba-Geigy that its antiplatelet drug, Anturane, was effective in preventing sudden-onset mortality during the first six months after myocardial infarction.<sup>147</sup> The results of the

<sup>138</sup>*Castaneda v. Partida*, 430 U.S. 482, 496 n.17 (1977).

<sup>139</sup>*Id.*

<sup>140</sup>Michael O. Finkelstein, *The Application of Statistical Decision Theory to the Jury Discrimination Cases*, 80 HARV. L. REV. 338, 357–59 (1966).

<sup>141</sup>*Id.* at 359–60.

<sup>142</sup>*Hazelwood School District v. United States*, 433 U.S. 299, 311 n.17 (1977). The majority and dissenting opinions in *Hazelwood* even differed on whether to use a one-sided or two-sided significance test. See Paul Meier et al., *What Happened in Hazelwood: Statistics, Employment Discrimination, and the 80% Rule*, in STATISTICS AND THE LAW 1, 14 (Morris H. DeGroot et al. eds., 1986).

<sup>143</sup>For more on the use of *p*-values and significance testing in the legal system, see Meier et al., *supra* note 142, at 6–15; DAVID L. FAIGMAN ET AL., SCIENCE IN THE LAW: STANDARDS, STATISTICS AND RESEARCH ISSUES 188–98 (2002).

<sup>144</sup>Temple, *supra* note 18, at 1655.

<sup>145</sup>*Id.*

<sup>146</sup>Louis Lasagna, *Congress, the FDA, and New Drug Development: Before and After 1962*, 32 PERSP. BIOLOGY & MED. 322, 334 (1989). See also CARPENTER, *supra* note 112, at 363 (describing the gradual submission of U.S. pharmaceutical companies to the new FDA protocols).

multicenter trial gave  $p$ -values of 0.058 for cardiac mortality at 24 months and 0.041 for reduction in sudden death over that time, both compared to placebo.<sup>148</sup> The same outcomes were assessed in the period of two through seven months after myocardial infarction as well, both resulting in  $p$ -values well below 0.05.<sup>149</sup> FDA rejected the claim and published a critique of Ciba-Geigy's results in the *New England Journal of Medicine* explaining the agency's reasoning. FDA reviewers objected to several design features of the study, most notably definitions of outcome events, post hoc exclusions of patients who died during the study due to non-cardiac events, and multiple comparisons. This led the reviewers to state that their "major criticisms of the study are not statistical."<sup>150</sup> Despite this, they calculated adjusted  $p$ -values accounting for the design flaws and multiple testing, and got values in the 0.12–0.20 range. They finally concluded that the trial was "an insufficient basis for FDA approval."<sup>151</sup> Rather than establishing any specific statistical rules, FDA reviewers demonstrated their commitment to overall principles of study design and control of Type I error. One of the reviewers, Robert Temple, later called the critique "a public tutorial on the analytic problems that could arise in an otherwise well-conducted study."<sup>152</sup> The influence of statistics at FDA had never been stronger.

### C. FDA Guidance and the Formalization of the Two-Trial, 0.05 Standard

Later that decade, FDA published guidance for industry specifying the statistical analyses that would be required for NDA approval. The 1988 guidelines required drug sponsors to detail their statistical analysis plans, including primary outcomes measured and comparisons made. Specifically, FDA requested the use of methods ensuring adequate power and Type I error control, demonstrating that FDA assumed a hypothesis testing procedure with a pre-specified alpha level would be used. The  $p$ -value is specifically mentioned as a desired feature of the statistical analysis, and the presumption is made that  $p$ -values would be two-sided unless otherwise specified.<sup>153</sup> While a presumed 0.05 alpha level is not explicitly stated, the presentation in sample tables of 95 percent confidence intervals, which generally coincide with hypothesis tests at the 0.05 alpha level, suggests that a two-sided significance level of 0.05 would be reasonable.<sup>154</sup>

In the 1990s, FDA sought to harmonize its guidance with that of similar agencies in other regions or countries, especially the European Union and Japan. In 1996, the International Conference on Harmonization issued Consolidated Guidance for Industry. The document focused primarily on the design of trials and ensuring ethical treatment of participants. The brief sections on efficacy determinations and statistics, though general, did include the principles of planning statistical tests with pre-specified significance levels and using power calculations to determine appropriate sample sizes.<sup>155</sup>

<sup>147</sup>Robert Temple & Gordon W. Pledger, *The FDA's Critique of the Anturane Reinfarction Trial*, 303 NEW ENGL. J. MED. 1488, 1488 (1980).

<sup>148</sup>Anturane Reinfarction Trial Research Group, *Sulfapyrazole in the Prevention of Sudden Death after Myocardial Infarction*, 302 NEW ENGL. J. MED. 250, 252–254 (1980).

<sup>149</sup>*Id.*

<sup>150</sup>Temple & Pledger, *supra* note 147, at 1492.

<sup>151</sup>*Id.*

<sup>152</sup>Temple, *supra* note 18, at 1656.

<sup>153</sup>U.S. FOOD & DRUG ADMIN., GUIDELINE FOR THE FORMAT AND CONTENT OF THE CLINICAL AND STATISTICAL SECTIONS OF AN APPLICATION 68–70 (1988).

<sup>154</sup>*Id.* at 107–08.

In 1997, the FDA Modernization Act (FDAMA) amended Section 505 of the FDCA. The law did not specifically adjust the standards for evidence, except in one area detailed *infra* section IV.D. The bigger changes served to encourage FDA to conduct reviews of NDAs more efficiently. For example, the statute mandated that FDA officials meet with drug sponsors “for the purpose of reaching agreement on the design and size of clinical trials intended to form the primary basis of an effectiveness claim.”<sup>156</sup> The agreed-upon parameters for the trial would be binding upon the sponsor and the reviewers. While no statistical requirements explicitly entered into the law, the principle of FDA and sponsors agreeing on levels at which to control Type I and Type II errors became codified in the statute.

The next year, FDA released guidance specifically discussing statistical questions in clinical trials. “Statistical Principles for Clinical Trials” did not detail specific procedures or methodologies, but laid out principles to guide the sponsor’s statisticians in conducting trial design and analysis. It does, however, specifically address the alpha level and power questions:

The treatment difference to be detected may be based on a judgement concerning the minimal effect which has clinical relevance in the management of patients or on a judgement concerning the anticipated effect of the new treatment, where this is larger. Conventionally, the probability of Type I error is set at 5 percent or less or as dictated by any adjustments made necessary for multiplicity considerations; the precise choice may be influenced by the prior plausibility of the hypothesis under test and the desired impact of the results. The probability of Type II error is conventionally set at 10 percent to 20 percent. It is in the sponsor’s interest to keep this figure as low as feasible, especially in the case of trials that are difficult or impossible to repeat. Alternative values to the conventional levels of Type I and Type II error may be acceptable or even preferable in some cases.<sup>157</sup>

FDA here explicitly suggested a 0.05 alpha level, but left the door open to other thresholds if justified. In addition, FDA suggested a focus on clinical relevance in determining whether an effect is significant. Later in the document, FDA endorses the general use of two-sided hypothesis tests, in order to match two-sided confidence intervals, unless there is reason to prefer one-sided tests.<sup>158</sup> The 1998 guidance represented the most explicit statement yet of the role of the 0.05 alpha level in FDA’s decision-making.

Since the 1998 guidance, two trials, each with a two-sided alpha level of 0.05, has remained the paradigm for FDA approval of drug efficacy. Performing multiple hypothesis tests (either for multiple endpoints, subgroup analyses, or for potentially stopping the trial before its scheduled end) can lead to adjustments of this level, but the goal is generally to ensure the Type I error rate of each trial is less than 0.05. In recent years, guidance has not been as explicit about the 0.05 standard, but a focus on controlling Type I error and summarizing

---

<sup>155</sup>U.S. FOOD & DRUG ADMIN./INTL. CONF. ON HARMONIZATION, GUIDANCE FOR INDUSTRY: E6 CLINICAL GOOD PRACTICE: CONSOLIDATED GUIDANCE 41–42 (1996).

<sup>156</sup>Pub. L. No. 105–115, 111 Stat. 2316 (1997) (codified at 21 U.S.C. § 355(b)(5)(B)(i)(I)).

<sup>157</sup>U.S. FOOD & DRUG ADMIN., GUIDANCE FOR INDUSTRY: E9 STATISTICAL PRINCIPLES FOR CLINICAL TRIALS 22 (1998).

<sup>158</sup>*Id.* at 32.

statistical evidence through  $p$ -values has continued.<sup>159</sup> This continued preference has been borne out in statements by FDA officials, approvals of drug submissions, and the academic literature in both law and biomedicine.<sup>160</sup>

The importance of this standard became clear very quickly, in an FDA decision on United Therapeutics' application for the drug Uniprost (later renamed Remodulin). Submitted in October 2000, the NDA specified an alpha level of 0.049 ("the traditional standard for two confirmatory studies with an adjustment" for one subgroup test).<sup>161</sup> The statistical reviewer found this standard reasonable, and when the  $p$ -values for the two studies came in at 0.0607 and 0.0550 the reviewer found "no justification for stretching beyond what was specified in the protocol" and was additionally unpersuaded by a  $p$ -value derived from pooling those studies that was above the threshold for a single study.<sup>162</sup> This review led to a letter by the Director of the Division of Cardio-Renal Drug Products urging the non-approval of Uniprost.<sup>163</sup> In the end, a re-submission focusing on the results of a surrogate endpoint in the studies was approved, conditional upon beginning an additional postmarketing trial.<sup>164</sup>

In addition to showing a relatively strict adherence to the 0.05 threshold, the Uniprost statistical reviewer also offers a more general defense of Type I error control. The reviewer notes FDA's twin goals of keeping ineffective drugs off the market while approving effective drugs and appeals. In language similar to that used by Neyman and Pearson nearly 70 years earlier, he defends the Type I and Type II error control and significance levels traditionally set by FDA as the most appropriate way to strike that balance.<sup>165</sup>

A similar case occurred a few years later with Dendreon's prostate cancer immunotherapy drug, Provenge. The initial submission in 2006 included two trials that had been conducted with a pre-specified significance level of 0.049 (again to adjust for other analyses) for the outcome of disease progression. One of the trials resulted in a  $p$ -value of 0.052 for this endpoint, which was noted as a failure to meet the primary endpoint by the statistical reviewer.<sup>166</sup> The sponsor attempted to present an efficacy evidence argument based on overall mortality in the trials, which had not been specified as a primary endpoint and did not have a pre-specified significance level in the protocol. Despite impressive  $p$ -values for these results from each trial, the reviewer noted that these were "post-hoc analyses" and thus it was "difficult to interpret hypothesis test results" for them.<sup>167</sup> He suggested non-approval for efficacy, claiming that "[t]he evidence is not substantial from a statistical perspective."

<sup>159</sup>U.S. FOOD & DRUG ADMIN., INTEGRATED SUMMARY OF EFFECTIVENESS: GUIDANCE FOR INDUSTRY 10 (2015).

<sup>160</sup>See Joseph W. Cormier, *Advancing FDA's Regulatory Science through Weight of Evidence Evaluations*, 28 J. CONTEMP. HEALTH L. & POL'Y 8–10 (2011); Darrow, *supra* note 22, at 2113–14; Russell Katz, *FDA: Evidentiary Standards for Drug Development and Approval*, 1 NEURORX 307, 310–11 (2004).

<sup>161</sup>*Statistical Review for Uniprost, Application Number: 21–272*, U.S. FOOD & DRUG ADMIN. 3 (2001), [https://www.accessdata.fda.gov/drugsatfda\\_docs/nda/2002/21-272\\_Remodulin\\_statr.pdf](https://www.accessdata.fda.gov/drugsatfda_docs/nda/2002/21-272_Remodulin_statr.pdf).

<sup>162</sup>*Id.* at 18.

<sup>163</sup>Letter from Director, Division of Cardio-Renal Drug Products, U.S. FOOD & DRUG ADMIN., to Director, Office of Drug Evaluation and Review, U.S. FOOD & DRUG ADMIN. (Mar. 9, 2001), [https://www.fda.gov/OHRMS/DOCKETS/ac/01/briefing/3775b1\\_01\\_Division%20Director's%20memo%20for%20NDA%2021-272.htm](https://www.fda.gov/OHRMS/DOCKETS/ac/01/briefing/3775b1_01_Division%20Director's%20memo%20for%20NDA%2021-272.htm).

<sup>164</sup>Letter from Robert Temple, U.S. FOOD & DRUG ADMIN., to Dean Bruce, United Therapeutics Corporation (Feb. 8, 2002), [https://www.accessdata.fda.gov/drugsatfda\\_docs/nda/2002/21-272\\_Remodulin\\_Approv.pdf](https://www.accessdata.fda.gov/drugsatfda_docs/nda/2002/21-272_Remodulin_Approv.pdf).

<sup>165</sup>*Statistical Review for Uniprost*, *supra* note 161, at 6–7. *Cf.* Neyman & Pearson, *supra* note 89, at 291, 296, 302–05.

<sup>166</sup>*Statistical Review for Provenge, BLA/Serial Number: 125197/0*, U.S. FOOD & DRUG ADMIN. 16, 52 (2007), <https://www.fda.gov/BiologicsBloodVaccines/CellularGeneTherapyProducts/ApprovedProducts/ucm210012.htm>.

<sup>167</sup>*Id.* at 52.



harkening back to the statutory language.<sup>168</sup> FDA did not approve Provenge at that time, in large part due to the statistical shortcomings and lack of Type I error control on the mortality claims.<sup>169</sup> In 2010, after the sponsor conducted another, larger trial with mortality as a pre-specified endpoint and achieved statistically significant results ( $p$ -value of 0.032), the drug was approved.<sup>170</sup>

In all, FDA guidance in the late 1990s made explicit the two-trial, 0.05 standard. Subsequent drug approval decisions reiterated that position. Guidance in May 1998 laid out several reasons for the preference for two Phase 3 trials: (1) reducing the risk of unanticipated, unavoidable biases in any given investigation; (2) reducing the statistical Type I error rate; (3) reducing the possibility of center- or population-specific results leading to wider indications; and (4) reducing the risk of (rare) fraudulent results leading to improper decision-making.<sup>171</sup> This mix of statistical and non-statistical reasons underpins FDA's continued preference for two-trial evidence of efficacy. This preference, however, is not immovable, as is seen through the two other primary routes discussed in that guidance.

#### D. The Single Trial with Corroborating Evidence Standard

The May 1998 guidance discussed FDA's flexibility in acting on specific cases while illustrating its main paradigms for approval. For decades, though, FDA officials had stated that the two-trial standard, while supported by statute, was not a strict rule, and that a single adequate and well-controlled study could qualify as "substantial evidence."<sup>172</sup> This principle was made explicit with the passage of the FDA Modernization Act in 1997. The statute amended section 505(d) of the FDCA to add that the Secretary of Health and Human Services (or his or her designee, generally the FDA Commissioner), could determine "based on relevant science, that data from one adequate and well-controlled clinical investigation and confirmatory evidence ... are sufficient to establish effectiveness" and could then use that one study as the basis for approving a drug under the "substantial evidence" standard.<sup>173</sup>

This standard of a single study with corroborating evidence was laid out explicitly in the May 1998 guidance. Various types of corroborating evidence are suggested that may be appropriate, but all of them require a study of the drug with very strong results.<sup>174</sup> FDA specifically stated that any attempt to gain approval of a drug without two trials would leave "little room for study imperfections or contradictory (nonsupporting) information."<sup>175</sup> The remainder of this section of the guidance appealed primarily to biological principles and feasibility rather than statistical arguments.

<sup>168</sup>Id.

<sup>169</sup>Letter from Ashok Batra, U.S. FOOD & DRUG ADMIN., to Elizabeth C. Smith, Dendreon Corp. 3 (May 8, 2007), <https://www.fda.gov/BiologicsBloodVaccines/CellularGeneTherapyProducts/ApprovedProducts/ucm210012.htm>.

<sup>170</sup>Statistical Review and Evaluation for Provenge, BLA/Serial Number: 125197/0, U.S. FOOD & DRUG ADMIN. 7, 35–36 (2010), <https://www.fda.gov/BiologicsBloodVaccines/CellularGeneTherapyProducts/ApprovedProducts/ucm210012.htm>; Letter from Mary A. Malarkey & Celia M. Witten, U.S. Food & Drug Admin., to Elizabeth C. Smith, Dendreon Corp. (Apr. 29, 2010), <https://www.fda.gov/BiologicsBloodVaccines/CellularGeneTherapyProducts/ApprovedProducts/ucm210012.htm>.

<sup>171</sup>U.S. FOOD & DRUG ADMIN., GUIDANCE FOR INDUSTRY: PROVIDING CLINICAL EVIDENCE OF EFFECTIVENESS FOR HUMAN DRUGS AND BIOLOGICAL PRODUCTS 4–5 (1998).

<sup>172</sup>HUTT ET AL., *supra* note 5, at 725–26.

<sup>173</sup>Pub. L. No. 105–115, 111 Stat. 2313 (1997) (codified at 21 U.S.C. § 355(d)).

<sup>174</sup>U.S. FOOD & DRUG ADMIN., *supra* note 171, at 8–12.

<sup>175</sup>Id. at 6.

## E. The Single Multi-Center Trial Standard

More compelling statistically, and perhaps more enticing to drug sponsors, was the guidance laid out for a single, multi-center trial standard. As clinical trials grew larger and more complex throughout the decades, the possibility of a single trial providing evidence as convincing as that from two trials grew as well. Prior to the guidance, FDA had made some approvals on the basis of large, single trials, especially if there were specific reasons a second trial would be unfeasible or unethical. In this guidance, FDA made that standard explicit, but still warned that two-trial conclusions are “more secure” than one-trial conclusions.<sup>176</sup>

In the guidance, FDA enumerated five characteristics that may allow a single study to suffice. The study must generally be a multicenter study, have subsets of the study population that show consistent results, have multiple pairwise comparisons showing an effect, and demonstrate an effect on several distinct, important outcomes or endpoints. The final characteristic is that the study shows a “[s]tatistically very persuasive finding.” This is interpreted to be a “very low *p*-value,” preferably accompanied by “very sizable treatment effects.” While no specific figures are given, the wording suggests that levels of evidence even greater than the traditional alpha level of 0.05, or even 0.01, may be necessary.<sup>177</sup>

The guidance cited two drugs that had succeeded in using this pathway already, timolol for preventing complications after myocardial infarction (MI) and combination streptokinase/aspirin for preventing mortality among patients with suspected MI. Both were cited in particular for their persuasive statistical results.<sup>178</sup> For timolol, the single trial involved 20 hospitals and 1,884 enrolled patients. The primary endpoint of mortality over 33 months after MI was tested and timolol showed a 39.4 percent reduction compared to placebo, with a *p*-value of 0.0003.<sup>179</sup> The streptokinase/aspirin trial (known as ISIS II) had 417 participating hospitals with 17,187 patients randomized into four arms: streptokinase alone, aspirin alone, combination of streptokinase and aspirin, and neither. The combination therapy group had a 42 percent reduction in vascular mortality compared to the placebo, with a *p*-value less than 0.00001. The combination therapy performed better than either therapy individually, with *p*-values less than 0.0001 for the comparison of combination therapy with each individual therapy.<sup>180</sup> In both cases, then, a single, large, multi-center study with substantial effect sizes and very low *p*-values obviated the need for a second confirmatory study in the eyes of FDA.

No concrete standard has been set for what constitutes a “very persuasive” *p*-value. The Uniprost decision discussed *supra* section IV.C referred to a 0.00125 standard as the traditional one-study guidance from the Division of Cardio-Renal Drugs.<sup>181</sup> A presentation on topical microbicides from FDA officials in 2003 suggested that one trial would need a *p*-

<sup>176</sup>*Id.* at 13.

<sup>177</sup>*Id.* at 13–15.

<sup>178</sup>*Id.* at 12, 15.

<sup>179</sup>Terje Pedersen, *The Norwegian Multicenter Study on Timolol after Myocardial Infarction - Design, Management and Results on Mortality*, 210 J. INTERNAL MED. 235 (1981).

<sup>180</sup>ISIS-2 (Second International Study of Infarct Survival) Collaborative Group, *Randomised Trial of Intravenous Streptokinase, Oral Aspirin, Both, or Neither among 17 187 Cases of Suspected Acute Myocardial Infarction: ISIS-2*, 332 LANCET 349, 354 (1988).

<sup>181</sup>*Statistical Review for Uniprost*, *supra* note 161, at 5–6.

value less than 0.001 to be considered. They justified this by noting that, absent non-statistical validity considerations, this ensures that the Type I error rate is no greater than the two-trial, 0.05 standard alpha level.<sup>182</sup> One year later, another FDA official, also discussing topical microbicides, stated that a single trial, significant at the 0.001 level, would be “persuasive, robust” evidence, while one significant at the 0.01 level would be “acceptable” if the study had other supportive data and good internal consistency.<sup>183</sup> A reviewer of a new drug in 2009 suggested that “less than 0.01” would likely be required and thus rejected “one study with a marginally significant p-value.”<sup>184</sup> In perhaps the most definitive statement, Robert Temple remarked in 2005 that “we ordinarily have said that a value in the neighborhood of 0.001 is good enough for a single trial.”<sup>185</sup>

The pharmaceutical company Nuvelo and FDA prospectively agreed to the stringent 0.00125 standard for a single trial for approval of the company’s thrombolytic agent altimeprase.<sup>186</sup> The trial, known as SONOMA-2, concluded in 2007. For the primary efficacy endpoint, dissolution of blood clots, the drug achieved a *p*-value of 0.022 against placebo.<sup>187</sup> This result, significant at the traditional 0.05 level but not at the one-trial 0.00125 level, led to Nuvelo withdrawing the NDA submission. Nuvelo’s shareholders sued the company, in part alleging that they were misled by the use of a 0.00125 significance level rather than the 0.05 level that has “traditionally been considered convincing evidence by the FDA.”<sup>188</sup> The case was settled prior to a decision on the merits by the District Court for the Northern District of California, so the court did not opine specifically on the reasonableness of the 0.00125 significance level.<sup>189</sup>

Notwithstanding some failures, drug sponsors have used the single-trial approval pathway with some frequency since the guidance document, and the two cases cited therein, raised the potential for such approval. A study of new drug approvals at FDA from 2005 to 2015 found that nearly 37 percent of new drugs were approved on the basis of one pivotal efficacy trial (note that this may include drugs that fell under the pathway described *supra* section IV.D as well as the purely single-trial standard). The proportion was highest among cancer drugs, with over 80 percent of new cancer drugs approved on the basis of only one trial.<sup>190</sup> In an address to Congress shortly before stepping down from her role, FDA Commissioner

<sup>182</sup>Rafia Bhore & Greg Soon, *Statistical Considerations for Topical Microbicide Phase 2 and 3 Trial Designs*, U.S. FOOD & DRUG ADMIN. 11 (Aug. 20, 2003), [https://www.fda.gov/ohrms/dockets/ac/03/slides/3970S1\\_07\\_Bhore.ppt](https://www.fda.gov/ohrms/dockets/ac/03/slides/3970S1_07_Bhore.ppt) [<https://perma.cc/Q4X2-8DY8>]. See also *Statistical Review for Uniprost*, *supra* note 161, at 7 (using the same reasoning in explaining a one-trial 0.00125 standard). See also Zhenming Shun et al., *Statistical Consideration of the Strategy for Demonstrating Clinical Evidence of Effectiveness—One Larger vs Two Smaller Pivotal Studies*, 24 STAT. MED. 1622–28 (2005) (detailing the underlying statistics for combining evidence from different studies and controlling overall Type I error).

<sup>183</sup>Teresa C. Wu, *Clinical Development of Topical Microbicides: U.S. Regulatory Perspective*, U.S. FOOD & DRUG ADMIN., 13 (2004), <https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/InvestigationalNewDrugINDApplication/Overview/UCM166921.pdf> [<https://perma.cc/AEE3-AFRF>].

<sup>184</sup>*Statistical Review and Evaluation for Bystolic, NDA/Serial Number: 22–742*, U.S. FOOD & DRUG ADMIN. 3 (2009), <https://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/CardiovascularandRenalDrugsAdvisoryCommittee/UCM196558.pdf> [<https://perma.cc/7GXV-24T7>].

<sup>185</sup>Temple, *supra* note 108.

<sup>186</sup>*In re Nuvelo, Inc. Securities Litigation* 688 F. Supp. 2d 1218, 1221 (N.D. Cal. 2009).

<sup>187</sup>Stephan Moll et al., *Safety and Efficacy of Altimeprase in Subjects with Occluded Central Venous Access Devices: The SONOMA-2 Study* 110(11) BLOOD 552A–53A (2007).

<sup>188</sup>*In re Nuvelo, Inc. Securities Litigation* 688 F. Supp. 2d 1218, 1221 (N.D. Cal. 2009).

<sup>189</sup>*In re Nuvelo, Inc. Securities Litigation* No. 3:07 (N.D. Cal. July 13, 2012) (order cancelling hearing on lead plaintiffs’ unopposed motion for distribution of settlement fund).

<sup>190</sup>Nicholas S. Downing et al., *Clinical Trial Evidence Supporting FDA Approval of Novel Therapeutic Agents, 2005–2012*, 311 J. AM. MED. ASS’N 375 (2014).



Margaret Hamburg, in defending the speed, rigor, and flexibility of the agency's approval process, noted similarly that one-third of new drugs were approved on the basis of single clinical trials.<sup>191</sup>

Between the FDA Modernization Act and the guidance documents of the late 1990s, FDA set down fairly clear expectations regarding statistical analyses to be included in drug approval submissions. While not an explicit or hard-and-fast rule, the 0.05 alpha standard has remained, generally implicitly, in these guidelines and has even given rise to a general 0.001 alpha standard for a single-trial approval. Fisher's appeal to a "convenient" and "customary" level in the 1920s and 1930s has thus survived over eight decades to inform policy today. But just as Fisher's hypothesis testing framework faced challenges in his own time, FDA's standards and the  $p$ -value as a whole have faced challenges in recent years.

#### IV. Statistics at FDA: Contemporary Challenges and Future Directions

Ever since the initial disagreements between Neyman and Pearson and Fisher, the  $p$ -value and the 0.05 alpha level standard have caused controversy. Statisticians have debated the proper use and interpretation of the  $p$ -value, to the point of questioning whether it belongs in statistical reasoning at all. Policymakers and FDA stakeholders have questioned whether the roles of the  $p$ -value and the 0.05 standard in drug approvals have been appropriate. And there has been no shortage of alternatives presented. Even as the field of statistics has changed and FDA's standards have adjusted to new trial designs and statistical analysis plans, the  $p$ -value has not lost its influence, or controversy, at the agency.

##### A. Challenges to the P-Value Paradigm

The challenge to Fisher's interpretations by Jerzy Neyman and Egon Pearson, described *supra* section III.D, was only the first challenge to the regime of significance testing via the  $p$ -value. In the 1960s, as FDA was incorporating significance testing into its new drug efficacy pre-approval regime, prominent psychologists pointedly questioned the appropriate role of the  $p$ -value in their field. In 1960, William Rozeboom wrote in the *Psychological Bulletin* of the failings of the  $p$ -value and the significance testing regime. While many of his arguments focused on his philosophical disdain for the idea of accepting or rejecting a scientific hypothesis outright, he also addressed the more mathematical question of the significance level to be used. "There is no reason (at least provided by the method)," Rozeboom wrote, "why the point of statistical 'significance' should be set at the 95% level, rather than, say the 94% or 96% level. Nor does the fact that we sometimes select a 99% level of significance, rather than the usual 95% level, mitigate this objection—one is as arbitrary as the other."<sup>192</sup> He comes back to this point later in the article, questioning "what scientist in his right mind would ever feel that there is an appreciable difference between the interpretative significance of data, say, for which one-tailed  $p = .04$  and that of data for which  $p = .06$ , even though the point of 'significance' has been set at  $p = .05$ ?"<sup>193</sup>

<sup>191</sup>Statement of Margaret A. Hamburg, U.S. FOOD & DRUG ADMIN., to Committee on Health, Education, Labor and Pensions, U.S. Sen. (Mar. 10, 2015), <https://www.fda.gov/newsevents/testimony/ucm437481.htm> [<https://perma.cc/MLP9-LA8J>].

<sup>192</sup>William W. Rozeboom, *The Fallacy of the Null-Hypothesis Significance Test*, 57 PSYCHOL. BULL. 423 (1960).

<sup>193</sup>*Id.* at 424.

Rozeboom suggests alternatives that have now become common in journals, and are becoming more common in FDA reviews: the confidence interval and Bayesian inference, both discussed in more detail *infra* section V.B.

Not long after Rozeboom, in his own words, “vigorously excoriated” the significance testing procedure,<sup>194</sup> other psychologists continued the argument. David Bakan, in the same publication in 1966, lamented “a kind of essential mindlessness in the conduct of research” when investigators focus, to the exclusion of other inference, on *p*-values and significance testing.<sup>195</sup> He agreed with Rozeboom’s prescription to focus more on effect sizes, confidence intervals, and Bayesian methods. In a far-reaching paper in *Philosophy of Science* in 1967, Paul Meehl contrasted the use of statistical tests in physics, wherein they serve to quantify the uncertainty of a numerical estimate, with their use in psychology, wherein they serve to accept or reject a null hypothesis.<sup>196</sup> All of these papers, along with others along the same lines, point out that a significance testing framework is more appropriate for a situation where a decision one way or another must be made than for the general principle of scientific inference. They also, however, question the model of significance testing more fundamentally, pointing out frequent misinterpretations of its results and the methodologically unjustified but “widespread adoption of the probabilities .01 or .05 as the allowable theoretical frequency of Type I errors” in the biological and social sciences.<sup>197</sup>

While psychologists and social scientists lamented the strict significance tests, they often faced different questions than biomedical researchers, especially those analyzing clinical trials. One sociologist wryly noted that “usually the only real decision facing a social scientist is whether to publish or suppress his findings,”<sup>198</sup> a sharp contrast with the clear-cut decision to approve or reject faced by FDA. In addition, biologists working under controlled conditions in randomized experiments have more control over design and thus may accept different thresholds and stricter significance tests than social scientists. The sociologist Sanford Labovitz wrote that “[u]nder such highly controlled conditions [of agricultural experiments] Fisher seemed justified in using the larger error rate of .05 instead of .01 or lower.”<sup>199</sup>

Nonetheless, in biology and medicine, critiques began to appear along similar lines as in the social sciences. As early as 1951, Fisher’s collaborator Frank Yates wrote of his concerns about the use of tests with strict levels of significance: “scientific workers have often regarded the execution of a test of significance on an experiment as the ultimate objective.”<sup>200</sup> In a speech delivered to the International Biometric Society in 1969, the outgoing British Regional President J.G. Skellam presented a defense of confidence intervals and some support for Bayesian ideas, warning that doctrinaire use of Fisherian significance

<sup>194</sup>*Id.* at 428.

<sup>195</sup>David Bakan, *The Test of Significance in Psychological Research*, 66 PSYCHOL. BULL. 436 (1966).

<sup>196</sup>Paul E. Meehl, *Theory-Testing in Psychology and Physics: A Methodological Paradox*, 34 PHIL. SCI. (1967).

<sup>197</sup>*Id.* at 106.

<sup>198</sup>Skipper et al., *supra* note 88, at 158.

<sup>199</sup>Sanford Labovitz, *Criteria for Selecting a Significance Level: A Note on the Sacredness of .05*, THE SIGNIFICANCE TEST CONTROVERSY—A READER, 166, 169 (Denton E. Morrison & Ramon E. Henkel eds., 1970).

<sup>200</sup>Frank Yates, *The Influence of Statistical Methods for Research Workers on the Development of the Science of Statistics*, 46 J. AM. STAT. ASS’N 33 (1951).

tests might “exercise their own unintentional brand of tyranny over other ways of thinking.”<sup>201</sup>

The objections grew in the 1980s as the leading medical journals began to call for more discussion of point estimates and confidence intervals, in addition to  $p$ -values and significance testing. In 1988, in promoting these approaches, the statistical adviser to the *British Heart Journal* provocatively titled his editorial “The end of the  $p$  value?”<sup>202</sup> Other prominent medical and epidemiologic researchers, including Kenneth Rothman, Richard Simon, and Steven Goodman, have also written along the same lines.<sup>203</sup> Rothman, in an editorial for *Annals of Internal Medicine* in 1986, wrote that “[t]esting for statistical significance continues today not on its merits as a methodological tool but on the momentum of tradition.”<sup>204</sup>

Common complaints about the  $p$ -value and significance testing have rested not only on methodological grounds, but also on the improper interpretation of results. The most common, and perhaps most reviled among statisticians, misinterpretation is the changing of the conditional probability.<sup>205</sup> In this falsehood, the  $p$ -value is taken to be the probability that the null hypothesis is true given the data. In fact, the  $p$ -value is the probability that the data would have occurred if the null hypothesis were true, and these two probabilities are very rarely the same. Another common error is ascribing clinical significance to a statistically significant result. A very large trial may achieve statistical significance at the 0.05 level even without any clinically meaningful difference in outcomes. And the overuse of testing procedures, without adjusting the significance level, or choosing tests based on seeing the data, known as “ $p$ -hacking,” has come under fire recently.<sup>206</sup>

The controversies about  $p$ -values came to a head in the 2010s, with some journals beginning to discourage or ban outright the use of the  $p$ -value in their pages.<sup>207</sup> In response to this, the American Statistical Association issued a wide-ranging statement on statistical significance and  $p$ -values in 2016.<sup>208</sup> With contributions and complementary articles by statisticians with a range of philosophical frameworks and preferred methodological techniques, the statement is hardly definitive or prescriptive. It does, however, go into great detail on proper use and interpretation of the  $p$ -value in the frequentist framework. Seeking relevance for statistical practice across scientific and policy-related disciplines, the statement emphasizes that “[w]hile the  $p$ -value can be a useful measure, it is commonly misused and misinterpreted.”<sup>209</sup> The statement concludes with a call for good study design and scientific inference, closing: “No single index should substitute for scientific reasoning.”<sup>210</sup>

<sup>201</sup>J. G. Skellam, *Models, Inference, and Strategy*, 25 BIOMETRICS 474 (1969).

<sup>202</sup>Stephen J. W. Evans et al., *The End of the P Value?*, 60 BRIT. HEART J. (1988).

<sup>203</sup>See, e.g., Steven N. Goodman, *A Comment on Replication, P-Values, and Evidence*, 11 STAT. MED. (1992); Goodman, *supra* note 35; Richard Simon, *Confidence Intervals for Reporting Results of Clinical Trials*, 105 ANNALS INTERNAL MED. (1986); Kenneth J. Rothman, *Significance Questing*, 105 ANNALS INTERNAL MED. (1986).

<sup>204</sup>Rothman, *supra* note 203, at 447.

<sup>205</sup>See, e.g., Goodman, *supra* note 35.

<sup>206</sup>See, e.g., Greenland et al., *supra* note 32, at 342–43, 346; Regina Nuzzo, *Statistical Errors*, 506 NATURE (2014).

<sup>207</sup>See, e.g., David Trafimow, *Editorial*, 36 BASIC & APPLIED SOC. PSYCHOL. (2014); David Trafimow & Michael Marks, *Editorial*, 37 BASIC & APPLIED SOC. PSYCHOL. (2015).

<sup>208</sup>Ronald L. Wasserstein & Nicole A. Lazar, *The ASA’s Statement on p-Values: Context, Process, and Purpose*, 70 AM. STAT. (2016).

<sup>209</sup>*Id.* at 131.

## B. Alternatives to the P-Value Paradigm

In light of these challenges to the  $p$ -value, many alternatives have been proposed. Remaining in the frequentist setting, a common call is for point estimates of effects and confidence intervals to be placed as prominently as  $p$ -values. Stuart Pocock's 1983 *Clinical Trials: A Practical Approach* encourages following significance tests with confidence limits (now more commonly called confidence intervals) "to estimate the magnitude of improvement of one treatment over another."<sup>211</sup> Frank Yates had previously endorsed this in 1951, lamenting that Fisher's works had "caused scientific research workers to pay undue attention to the results of tests of significance" and "too little to the estimates of the magnitude of the effects they are investigating."<sup>212</sup>

The confidence limits provide a range of estimates of the true effect under study such that, if the experiment is repeated under identical conditions many times, the given percentage of such confidence intervals will include the true parameter. As Pocock notes, "[i]t is standard practice to use 95% confidence limits."<sup>213</sup> This practice is as arbitrary, and largely derived from, the practice of using an alpha level of 0.05 in significance tests. Specifically, the link between confidence intervals and significance tests is clear: if the 95 percent confidence interval includes the value suggested by the null hypothesis, then the equivalent test will not be statistically significant at the 0.05 level. This feature of confidence intervals is sometimes, though not always, noted by its proponents. The use of the confidence interval brings the magnitude of effects and clinical relevance to the fore, but it does not do away with the problems of frequentist inference and does not lead to a clear decision rule that is different from the significance testing regime. It is worth noting that confidence limits were put forth, although not in their modern nomenclature, by both Ronald Fisher and Jerzy Neyman in the 1930s.<sup>214</sup>

For those who more fundamentally object to the frequentist paradigm, Bayesian statistics offers an attractive alternative. While this article cannot provide a reasonable treatment of the fundamentals of Bayesian inference, there is a benefit for clinical trial use that is fairly easy to understand. In many forms of Bayesian inference, an investigator begins with a prior belief about the parameter, or estimator, that she wishes to estimate (say, the difference in survival probabilities of those on a new treatment compared to those on a placebo). This prior belief is "updated" with information contained in the collected data through the use of the likelihood of that data appearing given the possible parameter values. Incorporating (formally, conditioning on) this data then gives an updated, or posterior, probability distribution for the parameter. This distribution can be described in many ways, with point estimates for the parameter given by the mean, median, or mode of the distribution and measures of uncertainty given by the standard deviation of the distribution or an interval in which the parameter has a certain probability of falling.<sup>215</sup> A Bayes Factor can also be calculated, which is the ratio of the posterior odds of one hypothesis and the prior odds of a

<sup>210</sup>*Id.* at 132.

<sup>211</sup>POCOCK, *supra* note 98, at 206.

<sup>212</sup>Yates, *supra* note 200, at 32. Fisher's student L. H. C. Tippett had also warned that the "statistical significance gives no information as to the magnitude or practical importance of any difference." TIPPETT, *supra* note 85, at 51.

<sup>213</sup>POCOCK, *supra* note 98, at 208.

<sup>214</sup>*See* DAVID & EDWARDS, *supra* note 100, at 189.

competing hypothesis (usually the alternative hypothesis compared to the null hypothesis).<sup>216</sup> This quantity has been referred to as a measure of the strength of the evidence contained in the data and is sometimes put forth as an alternative to the  $p$ -value.<sup>217</sup>

The Bayesian alternative provides intuitive results. The parameter does have a 95 percent probability of being within the 95 percent credible interval, unlike a frequentist 95 percent confidence interval, in the strictest sense. Since  $p$ -values are often mistakenly interpreted to indicate the probability of the parameter having some value given the data, the use of Bayesian results can be easier to explain.<sup>218</sup> Additionally, the use of a prior distribution for the parameter allows the investigator to incorporate past information into the model. This has led to some criticism of Bayesian inference, however, as it has been accused of being open to subjectivity on the part of the investigator in selecting the prior belief.<sup>219</sup> Because of this, some Bayesian proponents have proposed the use of non-informative priors, although this often leads to numerical results identical to those from frequentist inference. The use of sensitivity analyses, common for various reasons in statistical analysis of biomedical data, is often proposed to determine the influence of a prior.<sup>220</sup>

### C. FDA's Response to the Alternatives

FDA has encouraged the use of confidence intervals to represent trial results, but primarily as a supplement to testing procedures. In 1988, the Guideline for the Format and Content of the Clinical and Statistical Sections of an Application made some references to (usually 95 percent) confidence intervals as supplements to point estimates and its example tables include some confidence intervals around estimates. These do not take the place of  $p$ -values in reporting trial results, however, but supplement them.<sup>221</sup> The agency's 1998 Guidance for Industry on Statistical Principles for Clinical Trials gave fairly strong support to confidence intervals as supplements to significance tests and point estimates. "Estimates of treatment effects," the guidance states, "should be accompanied by confidence intervals, whenever possible."<sup>222</sup> In the 2015 guidance, the agency explicitly stated that sponsors should include estimated effect sizes, confidence intervals, and  $p$ -values in submissions: "A presentation of  $p$ -values alone would not be adequate."<sup>223</sup> Very narrow confidence intervals are mentioned as part of the statistical evidence that can lead to approval with one trial in the May 1998 guidance document as well.<sup>224</sup>

<sup>215</sup>The debate between frequentist probability and Bayesian inference has been going on for centuries, and the author has no designs on tackling the subject here; various books and articles explain both the fundamentals of Bayesian inference and the benefits and drawbacks of the Bayesian and frequentist frameworks. See, e.g., GELMAN ET AL., *supra* note 33, at 3–32 (covering in a fairly accessible manner the statistical basis for Bayesian inference); Ronald A. Fisher, *Inverse Probability*, 26 MATHEMATICAL PROC. CAMBRIDGE PHIL. SOC'Y (1930) (marking on the history of the Bayesian approach and critiquing that framework).

<sup>216</sup>GELMAN ET AL., *supra* note 33, at 184–86.

<sup>217</sup>Steven N. Goodman, *Toward Evidence-Based Medical Statistics. 2: The Bayes Factor*, 130 ANNALS INTERNAL MED. (1999).

<sup>218</sup>Duminda Wijesundera et al., *Bayesian Statistical Inference Enhances the Interpretation of Contemporary Randomized Controlled Trials*, 62 J. CLINICAL EPIDEMIOLOGY (2009).

<sup>219</sup>See, e.g., Goodman, *supra* note 217.

<sup>220</sup>*Id.*

<sup>221</sup>U.S. FOOD & DRUG ADMIN., *supra* note 153, at 39, 66, 68.

<sup>222</sup>U.S. FOOD & DRUG ADMIN., *supra* note 157, at 32.

<sup>223</sup>U.S. FOOD & DRUG ADMIN., *supra* note 159, at 8.

<sup>224</sup>U.S. FOOD & DRUG ADMIN., *supra* note 171, at 15.

Despite these supportive statements around confidence intervals, they have not supplanted  $p$ -values in significance determinations. This may be due in part to the fact that confidence intervals do not lend themselves to binary decision-making (except insofar as they are equivalent to hypothesis tests) and so are more appropriate for the building of evidence that occurs in medical literature, which generally “requires no firm decision.”<sup>225</sup> In addition, confidence intervals depend heavily on the size of an effect, which is largely why many practitioners prefer them. But FDA has generally interpreted its efficacy requirement to show “substantial evidence” that the drug has its purported effect, regardless of the magnitude of that effect, and distinctions between statistical significance and clinical significance are often blurred to the point of nonexistence.<sup>226</sup> Because of this, a  $p$ -value may be more appropriate for the evidentiary requirement than a confidence interval. Additionally, turning to 95 percent confidence intervals would not diminish the reliance on the 0.05 figure itself.

The Bayesian framework has generated considerable interest in the clinical trials field, especially over the last twenty years. Bayesian analyses of efficacy have been accepted for approval of drugs already (e.g., Pravigard Pac for prevention of myocardial infarction) and some trials are underway with prospective Bayesian designs.<sup>227</sup> In many ways, the medical devices field has led the way on the use of Bayesian trials. Certain classes of medical devices are subject to premarket approval by FDA, but the standard for proving effectiveness is not the same as for drugs and, in fact, “seems more flexible than the ‘substantial evidence’ standard applicable to drugs.”<sup>228</sup> Because of this, there has been more room for alternative statistical techniques, including Bayesian methods to gain ground.

In the drug sphere, the 1998 Statistical Principles for Clinical Trials noted that it focused on frequentist methods but that “[t]his should not be taken to imply that other approaches are not appropriate; the use of Bayesian ... and other approaches may be considered when the reasons for their use are clear and when the resulting conclusions are sufficiently robust.”<sup>229</sup> In 2004, FDA and Johns Hopkins University jointly held a conference entitled “Can Bayesian Approaches to Studying New Treatments Improve Regulatory Decision Making?”, which covered drugs as well as devices. The August 2005 issue of the journal *Clinical Trials* covered this workshop, publishing many of the talks given there.<sup>230</sup> In her remarks at the workshop, then-Acting Deputy Commissioner for Operations at FDA, Janet Woodcock, encouraged participants to “push forward in the Bayesian area.”<sup>231</sup> Robert Temple, then the Director of the Office of Medical Policy at CDER, took a somewhat more cautious tone. While expounding on the ways in which prior information was incorporated into FDA approval processes, including by lowering the  $p$ -value standards that might be required for a

<sup>225</sup>Jonathan A. C. Sterne & George D. Smith, *Sifting the Evidence—What’s Wrong with Significance Tests?* 322 BRIT. MED. J. 226, 229 (2001).

<sup>226</sup>Darrow, *supra* note 22, at 2125–26.

<sup>227</sup>See, e.g., Donald A. Berry, *Bayesian Clinical Trials*, 5 NATURE REV. DRUG DISCOVERY 27, 27 (2006); Donald A. Berry, *Adaptive Clinical Trials: The Promise and the Caution*, 29 J. CLINICAL ONCOLOGY 606, 606–07 (2011) [hereinafter Berry, *Adaptive*].

<sup>228</sup>HUTT ET AL., *supra* note 5, at 1236.

<sup>229</sup>U.S. FOOD & DRUG ADMIN., *supra* note 157, at 4.

<sup>230</sup>Norris E. Alderson, *Introduction*, 2 CLINICAL TRIALS 271, 271 (2005).

<sup>231</sup>Janet Woodcock, *FDA Introductory Comments: Clinical Studies Design and Evaluation Issues*, 2 CLINICAL TRIALS 273, 275 (2005).



single trial, he noted that explicit Bayesian proposals were still very rare for drug trials. And he warned against generalizing from the experience of device approvals, noting that device manufacturers may be “much more prepared to make assumptions about what to expect. It is therefore not really clear that the CDRH (Center for Devices and Radiation Health) studies and drug studies are exactly the same in that sense.”<sup>232</sup> Despite promise in the device field, then, and a workshop and an issue of a prominent journal devoted to explicating Bayesian trial approaches, the prospects were not necessarily bright for a wholesale renovation of the drug approval architecture.

Since then, there has been some uptake of Bayesian designs, but it remains limited. The biggest growth has been in the use of Bayesian methods in adaptive clinical trials. As adaptive designs have grown to get the most information out of a limited number of trial participants, the Bayesian analysis methods that conform well to updating with additional data have been frequently (but not always) paired with these designs.<sup>233</sup>

Draft guidance issued in February 2010 entitled “Adaptive Design Clinical Trials for Drugs and Biologics” gave strong indications that FDA was open to adaptive designs and expected to work with sponsors to craft design and analysis plans that could lead to approval. But while it acknowledged the value of Bayesian analysis methods in these designs, it fell back to the old standard with regards to alpha levels: “In general, the study design should be planned in a frequentist framework to control the overall study Type I error rate.”<sup>234</sup> While Bayesian methods were acknowledged, FDA was not ready to part with the significance testing framework that had defined drug approvals for decades. To date, this draft guidance has not been finalized; finalized guidance on adaptive designs for device trials issued in July 2016 was slightly more favorable to Bayesian methods but also recommended controlled Type I error rate.<sup>235</sup>

While statisticians, biomedical investigators, and journal authors in related fields have debated the proper role of  $p$ -values, significance testing, and the entire frequentist framework, these techniques have largely remained the law of the land for FDA. Several new approaches have been used to some degree, but none have led to a substantial reduction in the reliance on  $p$ -values and Type I error control. This has not stopped academics and trialists from proposing further refinements, however. Some, like the “split-sample analysis” process proposed by Mark van der Laan and coauthors, transform how error is controlled but remain rooted in the significance testing framework and appeal to overall 0.05 alpha levels.<sup>236</sup> This process uses some trial data as an exploratory set to identify subgroups on which the drug may be safe and efficacious. The remaining data are used to confirm safety and efficacy in these subgroups, with a higher-than-usual statistical significance standard applied to control the overall Type I error rate.<sup>237</sup> Others, like the Bayesian Decision Analysis

<sup>232</sup>Temple, *supra* note 108, at 281.

<sup>233</sup>Berry, *Adaptive*, *supra* note 227, at 606. *See also* Shein-Chung Chow & Mark Chang, *Adaptive Design Methods in Clinical Trials—A Review*, 3 ORPHANET J. RARE DISEASES 1, 1–2 (2008).

<sup>234</sup>U.S. FOOD & DRUG ADMIN., GUIDANCE FOR INDUSTRY: ADAPTIVE DESIGN CLINICAL TRIALS FOR DRUGS AND BIOLOGICS 34 (2010).

<sup>235</sup>U.S. FOOD & DRUG ADMIN., ADAPTIVE DESIGNS FOR MEDICAL DEVICE CLINICAL STUDIES: GUIDANCE FOR INDUSTRY AND FOOD AND DRUG ADMINISTRATION STAFF 14 (2016).

<sup>236</sup>Mark Van der Laan et al., *Improving the FDA Approval Process* at 5 (John M. Olin Program in Law and Economics Working Paper No. 580, 2011), [http://chicagounbound.uchicago.edu/law\\_and\\_economics/256/](http://chicagounbound.uchicago.edu/law_and_economics/256/) [<https://perma.cc/5G9H-U2PS>].

framework proposed by Leah Isakov and co-authors, explicitly break from past notions of error control and focus on some external standard for appropriate decision-making mechanisms.<sup>238</sup> This proposal assigns a cost to making an incorrect decision in the drug approval process based on the burden and severity of the disease in question and the safety and efficacy profile of the drug. A Bayesian framework is then used to determine the appropriate drug-specific threshold for the efficacy significance level for approval.<sup>239</sup> Given the reluctance to break from established (and well-understood) standards, however, it seems unlikely that any of these will see considerable use in the near future.

#### D. Patient Advocacy and Challenges to the FDA Regulatory Paradigm

In addition to the technical and statistical challenges, patients and their advocates, as well as industry-related voices, have challenged FDA's statistics-based approach to drug regulation. These objections reflect the tension between FDA's mandate to ensure the safety and efficacy of drugs sold in the United States and the goal of patients, their advocates, and the companies manufacturing and selling pharmaceuticals to ensure that therapies move from the lab to the consumer as quickly as possible and that scientists and companies are incentivized to generate new therapies.<sup>240</sup>

This opposition has a long history, and often depends on very specific circumstances or specific drugs. At the time of the passage of the Kefauver-Harris Amendments, the medical profession opposed the new powers bestowed upon FDA, the drug industry "acquiesced reluctantly," and "organized consumer groups heartily endorsed it."<sup>241</sup> As soon as FDA began putting regulations into effect, however, patients and doctors responded. When the Drug Efficacy Study recommended the removal of bioflavonoids from the market, consumers and their doctors began an intensive, though ultimately unsuccessful, lobbying effort. Foreshadowing many future arguments, one patient wrote to his senator about his concern at the regulators "countermand[ing] instructions of my personal physician of almost twenty years."<sup>242</sup>

This calculus has shifted repeatedly over time, but there are now very strong patient advocacy groups that oppose FDA decisions limiting access to new medicines. Health advocacy organizations exist for nearly every disease and medical condition, and can vary drastically in size and scope. Their main activities include the promotion of medical research, conducting disease awareness campaigns, and advocating "for policies that they believe are in their members' best interests."<sup>243</sup> Advocacy around clinical trials and FDA regulation increased drastically during the AIDS crisis, when patients and their supporters

<sup>237</sup>Id.

<sup>238</sup>Vahid Montazerhodjat & Andrew W. Lo, *Is the FDA Too Conservative or Too Aggressive?: A Bayesian Decision Analysis of Clinical Trial Design* 1, 1–2 (NBER Working Paper No. 21499, 2015), <http://www.nber.org/papers/w21499> [<https://perma.cc/8R38-YZZ8>].

<sup>239</sup>Id. at 14.

<sup>240</sup>See, e.g., Kulynych, *supra* note 22, at 127–28 (detailing the debate over this tension during debate over the 1997 FDA Modernization Act).

<sup>241</sup>Irving H. Jurow, *The Effect on the Pharmaceutical Industry of the "Effectiveness" Provisions of the 1962 Drug Amendments*, 19 FOOD DRUG COSM. L.J. 110, 112 (1964).

<sup>242</sup>CARPENTER, *supra* note 112, at 350.

<sup>243</sup>Sheila M. Rothman et al., *Health Advocacy Organizations and the Pharmaceutical Industry: An Analysis of Disclosure Practices*, 101 AM. J. PUB. HEALTH 602, 602 (2011).



aggressively challenged FDA to speed up drug approvals and expand the rights of patients to try experimental therapies.<sup>244</sup> As activities around drug access increased, some advocacy organizations began to partner with pharmaceutical companies, including by accepting funding from the corporations.<sup>245</sup> Other organizations advocated for agendas similar to those of the companies without forming explicit partnerships.<sup>246</sup> With aligned interests in speeding the time to market of new therapies, patient advocates and industry representatives both took issue with some of the statistical approaches undertaken by FDA, including the use of significance tests.

In 1987, FDA faced one of the first of a new class of large-molecule biological products, drugs derived from—or synthesized to replicate—complex natural substances whose structure cannot be readily determined, when biotechnology company Genentech submitted an application for approval of tissue plasminogen activator (TPA). In May, FDA refused to approve the TPA submission, instead requesting more data on drug efficacy and safety.<sup>247</sup> The advisory panel recommended this step in large part because of the statistical reviewer's determination that Genentech's studies failed to show a "measurable, beneficial effect" of the drug.<sup>248</sup> This decision was met with derision from the scientific and lay media.<sup>249</sup> The *Wall Street Journal* editorial page led the charge with the most emotional attacks on FDA's decision and FDA official Robert Temple himself. "Medical research has allowed statistics to become the supreme judge of its inventions," wrote the editors, who went on to ask, "Are American doctors going to let people die to satisfy the bureau of drugs' chi-square studies?"<sup>250</sup> Although FDA approved the drug later that year when presented with further study results,<sup>251</sup> the TPA question demonstrated the intense controversies that accompany any FDA drug rejection on efficacy grounds.

Even after the 1997 FDA Modernization Act helped speed up and expand access to therapies,<sup>252</sup> the debate over the proper role of FDA as gatekeeper to therapies has not abated. Controversies over drug approvals (or, more commonly, non-approvals) have continued, often led by patient advocacy groups. In particular, these advocacy groups raise concerns about Type II errors in drug decisions, FDA failing to approve a drug when it does in fact have an effect. Advocates, especially in the context of severe diseases with limited treatment options, point to unnecessary disease burdens and death because of these Type II errors.<sup>253</sup> As discussed *supra* section III.A, reducing Type II errors in statistical analyses would necessarily mean increasing the risk of Type I errors by raising the alpha level and approving more drugs that may be ineffective.

<sup>244</sup> *Id.* at 603.

<sup>245</sup> *See, e.g., id.* at 606–07.

<sup>246</sup> *See, e.g.,* CARPENTER, *supra* note 112, at 393–461.

<sup>247</sup> CARPENTER, *supra* note 112, at 2–5.

<sup>248</sup> Peter R. Kowey et al., *The TPA Controversy and the Drug Approval Process: The View of the Cardiovascular and Renal Drugs Advisory Committee*, 260 J. AM. MED. ASS'N 2250, 2251 (1988).

<sup>249</sup> *See, e.g.,* Daniel E. Koshland, Jr., *TPA and PDQ*, 237 SCIENCE 341, 341 (1987). *See also* CARPENTER, *supra* note 112, at 4 (describing several other critical articles, editorials, and statements by scientists).

<sup>250</sup> Editorial, *Human Sacrifice*, WALL ST. J., Jun. 2, 1987, at 30.

<sup>251</sup> Kowey et al., *supra* note 248, at 2252.

<sup>252</sup> Kulynych, *supra* note 22, at 127.

<sup>253</sup> Daniel P. Carpenter, *The Political Economy of FDA Drug Review: Processing, Politics, and Lessons for Policy*, 23 HEALTH AFFS. 52, 57–58 (2004).

The strict adherence to a 0.05 significance level may be unpalatable to advocates of particular treatments because of the risk of Type II errors. On the other hand, given its statutory mandate to ensure that drugs are approved only with “substantial evidence that the drug will have the effect it purports or is represented to have,” FDA must create some standard by which to assess the statistical evidence provided by clinical trials. With the biomedical community having come to accept hypothesis testing in the 1940s and 1950s, and Fisher’s 0.05 level, though arbitrary, becoming commonly used, any other level would be difficult to defend. Additionally, drug sponsors generally crave consistency and some foreknowledge that their clinical trial plan will lead to a positive result if the statistics meet the agreed-upon level, leading them to take FDA guidance very seriously and interact frequently with FDA officials.<sup>254</sup> To change the standard now would disrupt not only FDA’s processes for reviewing clinical trial data, but also sponsors’ processes for planning and analyzing trials, and public confidence in FDA’s past and future drug decisions. Any standard that did not rely on a specific significance level threshold would be open to challenges of subjectivity or inconsistent application, creating a much more difficult approval scheme for FDA to defend and justify statutorily.

## Conclusion

Over fifty years after the Kefauver-Harris Amendments created the drug efficacy review regime and nearly twenty years after the last major statutory change came with FDAMA, Section 505 was amended again. The 21st Century Cures Act, (Cures Act), an omnibus biomedical research bill passed in December 2016, introduced a number of novel ideas around the “substantial evidence” standard. The largest is the new Section 505(f), entitled “Real World Evidence,” which directs the Secretary of Health and Human Services to “evaluate the potential use of real world evidence” in drug approval processes and post-approval surveillance.<sup>255</sup> Real world evidence is then defined as “data regarding the usage, or the potential benefits or risks, of a drug derived from sources other than randomized clinical trials.”<sup>256</sup> A few paragraphs later, however, the bill clarifies that it “shall not be construed to alter ... the standards of evidence under ... section 505, including the substantial evidence standard in such subsection (d).”<sup>257</sup>

As the bill was nearing passage, then-FDA Commissioner Robert Califf and colleagues wrote in the *New England Journal of Medicine* about the promises and perils of “real world evidence.” They urged “caution” and tempering of “expectations of ‘quick wins,’” stating that “[r]eal-world research and the concepts of a planned intervention and randomization are entirely compatible.”<sup>258</sup> With no regulations or guidance documents yet promulgated expounding on the “real world evidence” mandate, it is unclear exactly what effect the new legislation will have on the drug approval standards. But some health policy researchers have warned that the legislation may “encourage use of less rigorous data to meet standards for

---

<sup>254</sup>Temple, *supra* note 18, at 1646–47.

<sup>255</sup>Pub. L. No. 114–255, 34 H.R. 64 (2016) (codified at 21 U.S.C. §355(f)).

<sup>256</sup>*Id.*

<sup>257</sup>*Id.*

<sup>258</sup>Rachel E. Sherman et al., *Real-World Evidence—What Is It and What Can It Tell Us?*, 375 NEW ENGL. J. MED. 2293, 2296 (2016).

approval.”<sup>259</sup> Others further suggest that new standards are already starting to influence FDA approvals, citing Sarepta Therapeutics’ Duchenne’s Muscular Dystrophy drug as “the accelerated approval of a drug with inadequate clinical trials and weak efficacy data.”<sup>260</sup>

In 2017, Portola Therapeutics decided to test the willingness of the new presidential administration and its FDA leadership to be flexible with statistical cutoffs. The drug in question, betrixaban, showed promise in early-stage trials in preventing blood clots after an illness. In its pivotal phase 3 trial, however, betrixaban failed to meet the pre-specified 0.05 alpha level for the primary endpoint, showing a  $p$ -value of 0.054.<sup>261</sup> Portola pointed FDA towards an alternative interpretation of the results, however, and won approval in late June.<sup>262</sup>

While the Portola case may indicate some changing standards at FDA, the further effects of the Cures Act and new FDA Commissioner Scott Gottlieb remain to be seen. But the use of  $p$ -values, significance testing, and the 0.05 alpha level at FDA have fifty years of history. They have survived in-fighting among the pioneers of the statistical methods, accusations of arbitrary cutoffs, a push for Bayesian statistics, a rejection of  $p$ -values in some disciplines and reassessment of their use in others, patient, physician, and pharmaceutical company challenges, and countless administrations and FDA officials. The  $p$ -value itself has been used to assess evidence for three hundred years and is intimately associated with the rise of the modern randomized clinical trial; the 0.05 significance level came with the  $p$ -value into the biomedical world from Fisher’s works on the subject. Adapted and re-assessed throughout the decades to meet the changing needs of FDA and respond to statistical and biomedical advances, this framework remains a cornerstone of U.S. pharmaceutical policy and looks poised to remain so for the foreseeable future.

“Almost every phase of the practice of medicine necessitates at least the rudimentary application of statistical ideas.”<sup>263</sup> As true as that statement by sociologist and pioneering National Cancer Institute epidemiologist Harold Dorn was in 1955, on the eve of FDA’s turn to statistical evidence, it is even more true today.

<sup>259</sup>Aaron S. Kesselheim & Jerry Avorn, *New “21st Century Cures” Legislation: Speed and Ease vs Science*, 317 J. AM. MED. ASS’N 581, 582 (2017).

<sup>260</sup>Amitabh Chandra & Rachel E. Sachs, *An FDA Commissioner for the 21st Century*, 376 NEW ENGL. J. MED. e31(1), e31(2) (2017).

<sup>261</sup>Alexander T. Cohen et al., *Extended Thromboprophylaxis with Betrixaban in Acutely Ill Medical Patients*, 375 NEW ENGL. J. MED. 534, 534 (2017).

<sup>262</sup>Letter from Dr. Richard Pazdur, U.S. FOOD & DRUG ADMIN., to Janice Castillo, Portola Pharmaceuticals, Inc. (Jun. 23, 2017), [https://www.accessdata.fda.gov/drugsatfda\\_docs/appletter/2017/208383Orig1s000ltr.pdf](https://www.accessdata.fda.gov/drugsatfda_docs/appletter/2017/208383Orig1s000ltr.pdf) [<https://perma.cc/J6B6-RMCR>]. See also Adam Feuerstein, *Will Trump’s FDA Relax Standards for Drug Approval? We’ll Get a Clue Soon with Decision on Portola’s Anticoagulant*, STAT NEWS (Jun. 21, 2017), <https://www.statnews.com/2017/06/21/portola-drug-fda-decision/>; Damian Garde & Adam Feuerstein, *Surprise Approval for Portola’s Blood Thinner has Biotech Eyeing a New Dawn at the FDA*, STAT NEWS (Jun. 23, 2017), <https://www.statnews.com/2017/06/23/portola-fda-approval/> (describing the evidence accumulated by Portola, the controversies surrounding the approval process, and potential future ramifications of the approval decision).

<sup>263</sup>Harold F. Dorn, *Some Applications of Biometry in the Collection and Evaluation of Medical Data*, 1(6) J. CHRONIC DISEASES 638, 638 (1955).

# Exhibit 57



## Continuous Auctions and Insider Trading

Albert S. Kyle

*Econometrica*, Vol. 53, No. 6. (Nov., 1985), pp. 1315-1336.

Stable URL:

<http://links.jstor.org/sici?sici=0012-9682%28198511%2953%3A6%3C1315%3ACAAIT%3E2.0.CO%3B2-8>

*Econometrica* is currently published by The Econometric Society.

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/econosoc.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

---

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

## CONTINUOUS AUCTIONS AND INSIDER TRADING

BY ALBERT S. KYLE<sup>1</sup>

A dynamic model of insider trading with sequential auctions, structured to resemble a sequential equilibrium, is used to examine the informational content of prices, the liquidity characteristics of a speculative market, and the value of private information to an insider. The model has three kinds of traders: a single risk neutral insider, random noise traders, and competitive risk neutral market makers. The insider makes positive profits by exploiting his monopoly power optimally in a dynamic context, where noise trading provides camouflage which conceals his trading from market makers. As the time interval between auctions goes to zero, a limiting model of continuous trading is obtained. In this equilibrium, prices follow Brownian motion, the depth of the market is constant over time, and all private information is incorporated into prices by the end of trading.

## 1. INTRODUCTION

HOW QUICKLY IS NEW PRIVATE INFORMATION about the underlying value of a speculative commodity incorporated into market prices? How valuable is private information to an insider? How does noise trading affect the volatility of prices? What determines the liquidity of a speculative market? The purpose of this paper is to show how answers to questions like these can be obtained as derived results by modelling rigorously the trading strategy of an insider in a dynamic model of efficient price formation.

In the particular model we investigate, one risky asset is exchanged for a riskless asset among three kinds of traders: a single insider who has unique access to a private observation of the *ex post* liquidation value of the risky asset; uninformed noise traders who trade randomly; and market makers who set prices efficiently (in the semi-strong sense) conditional on information they have about the quantities traded by others. Trading is modelled as a sequence of many auctions, structured to give the model the flavor of a sequential equilibrium as described by Kreps and Wilson [4].

At each auction trading takes place in two steps. In step one, the insider and the noise traders simultaneously choose the quantities they will trade (in effect, placing "market orders"). When making this choice, the insider's information consists of his private observation of the liquidation value of the asset, as well as past prices and past quantities traded by himself. He does not observe current or future prices, or current or future quantities traded by noise traders. The random quantity traded by noise traders is distributed independently from present or past quantities traded by the insider and independently from past quantities traded by noise traders. In step two, the market makers set a price, and trade the quantity which makes markets clear. When doing so, their information consists of observations of the current and past aggregate quantities traded by the insider and noise traders combined. We call these aggregate quantities the "order flow."

<sup>1</sup> The author thanks the Centre of Policy Studies, Monash University, and the Yale School of Organization and Management, Yale University, for their hospitality and support when this research was undertaken. The author also thanks Paul Milgrom, Peter Hartley, Phil Dybvig, and Avinash Dixit for useful comments.

Market makers do not observe the individual quantities traded by the insider or noise traders separately, nor do they have any other kind of special information. As a consequence, price fluctuations are always a consequence of order flow innovations.

The informed trader, who is risk neutral, is assumed to maximize expected profits. He acts as an intertemporal monopolist in the asset market, taking into account explicitly the effect his trading at one auction has on the price at that auction and the trading opportunities available at future auctions. The prices determined by market makers are assumed to equal the expectation of the liquidation value of the commodity, conditional on the market makers' information sets at the dates the prices are determined. Thus, market makers earn on average zero profits. Since they cannot distinguish the trading of the insider from the trading of noise traders, the noise traders in effect provide camouflage which enables the insider to make profits at their expense.

By assuming that the relevant random variables are normally distributed, the model acquires a tractable linear structure. This makes it possible to characterize explicitly a unique "sequential auction equilibrium" in which prices and quantities are simple linear functions of the observations defining the relevant information sets. In the limit as the time interval between auctions goes to zero, the discrete-time equilibrium converges to a particularly simple limit which we call a "continuous auction equilibrium." This equilibrium corresponds to what one obtains when the model is set up heuristically in continuous time.

In both the discrete model and the continuous limit, answers to the questions posed at the beginning of this paper are readily obtained. The informed trader trades in such a way that his private information is incorporated into prices gradually. In the continuous auction equilibrium where the quantity traded by noise traders follows a Brownian motion process, prices also follow Brownian motion. The constant volatility reflects the fact that information is incorporated into prices at a constant rate. Furthermore, all of the insider's private information is incorporated into prices by the end of trading in a continuous auction equilibrium. An *ex ante* doubling of the quantities traded by noise traders induces the insider and market makers to double the quantities they trade, but has no effect on prices, and thus doubles the profits of the insider.

Perhaps the most interesting properties concern the liquidity characteristics of the market in a continuous auction equilibrium. "Market liquidity" is a slippery and elusive concept, in part because it encompasses a number of transactional properties of markets. These include "tightness" (the cost of turning around a position over a short period of time), "depth" (the size of an order flow innovation required to change prices a given amount), and "resiliency" (the speed with which prices recover from a random, uninformative shock). Black [2] describes intuitively a liquid market in the following manner:

"The market for a stock is liquid if the following conditions hold:

- (1) There are always bid and asked prices for the investor who wants to buy or sell small amounts of stock immediately.
- (2) The difference between the bid and asked prices (the spread) is always small.



(3) An investor who is buying or selling a large amount of stock, in the absence of special information, can expect to do so over a long period of time at a price not very different, on average, from the current market price.

(4) An investor can buy or sell a large block of stock immediately, but at a premium or discount that depends on the size of the block. The larger the block, the larger the premium or discount.

In other words, a liquid market is a continuous market, in the sense that almost any amount of stock can be bought or sold immediately, and an efficient market, in the sense that small amounts of stock can always be bought and sold very near the current market price, and in the sense that large amounts can be bought or sold over long periods of time at prices that, on average, are very near the current market price."

Roughly speaking, Black defines a liquid market as one which is almost infinitely tight, which is not infinitely deep, and which is resilient enough so that prices eventually tend to their underlying value.

Our continuous auction equilibrium has exactly the characteristics described by Black. Furthermore, these aspects of market liquidity acquire a new prominence in our model because the insider, who does not trade as a perfect competitor, must make rational conjectures about tightness, depth, and resiliency in choosing his optimal quantity to trade. Moreover, depth and resiliency are themselves endogenous consequences of the presence of the insider and noise traders in the market. Market depth is proportional to the amount of noise trading and inversely proportional to the amount of private information (in the sense of an error variance) which has not yet been incorporated into prices. This makes our model a rigorous version of the intuitive story told by Bagehot [1]. Furthermore, our emphasis on the dynamic optimizing behavior of the insider distinguishes our model from the one of Glosten and Milgrom [3].

The plan of the rest of this paper is as follows: In Section 2, a single auction equilibrium is discussed in order to motivate the dynamic models which follow. In Section 3, a sequential auction equilibrium is defined, an existence and uniqueness result is proved, and properties of the equilibrium are derived. In Section 4, a continuous auction equilibrium is discussed heuristically, and in Section 5, it is shown that the continuous auction equilibrium is the limit of the sequential auction equilibrium as the time interval between auctions goes to zero. Section 6 makes some concluding comments.

## 2. A SINGLE AUCTION EQUILIBRIUM

In this section we motivate our equilibrium concept by discussing a simple model of one-shot trading.

*Structure and Notation.* The *ex post* liquidation value of the risky asset, denoted  $\tilde{v}$ , is normally distributed with mean  $p_0$  and variance  $\Sigma_0$ . The quantity traded by noise traders, denoted  $\tilde{u}$ , is normally distributed with mean zero and variance  $\sigma_u^2$ . The random variables  $\tilde{v}$  and  $\tilde{u}$  are independently distributed. The quantity traded by the insider is denoted  $\tilde{x}$  and the price is denoted  $\tilde{p}$ .

Trading is structured in two steps as follows: In step one, the exogenous values of  $\tilde{v}$  and  $\tilde{u}$  are realized and the insider chooses the quantity he trades  $\tilde{x}$ . When



doing so, he observes  $\tilde{v}$  but not  $\tilde{u}$ . To accommodate mixed strategies, the insider's trading strategy, denoted  $X$ , assigns to outcomes of  $\tilde{v}$  probability distributions defined over quantities traded. Since, however, mixed strategies are not optimal in what follows, the more intuitive interpretation of  $X$  as a measurable function such that  $\tilde{x} = X(\tilde{v})$  is justified. In step two, the market makers determine the price  $\tilde{p}$  at which they trade the quantity necessary to clear the market. When doing so they observe  $\tilde{x} + \tilde{u}$  but not  $\tilde{x}$  or  $\tilde{u}$  (or  $\tilde{v}$ ) separately. While their pricing rule, denoted  $P$ , can be defined to accommodate randomization, an intuitive interpretation of  $P$  as a measurable real function such that  $\tilde{p} = P(\tilde{x} + \tilde{u})$  is also justified.

The profits of the informed trader, denoted  $\tilde{\pi}$ , are given by  $\tilde{\pi} = (\tilde{v} - \tilde{p})\tilde{x}$ . To emphasize the dependence of  $\tilde{\pi}$  and  $\tilde{p}$  on  $X$  and  $P$ , we write  $\tilde{\pi} = \tilde{\pi}(X, P)$ ,  $\tilde{p} = \tilde{p}(X, P)$ .

*Definition of Equilibrium.* An *equilibrium* is defined as a pair  $X, P$  such that the following two conditions hold:

(1) *Profit Maximization:* For any alternate trading strategy  $X'$  and for any  $v$ ,

$$(2.1) \quad E\{\tilde{\pi}(X, P) | \tilde{v} = v\} \geq E\{\tilde{\pi}(X', P) | \tilde{v} = v\}.$$

(2) *Market Efficiency:* The random variable  $\tilde{p}$  satisfies

$$(2.2) \quad \tilde{p}(X, P) = E\{\tilde{v} | \tilde{x} + \tilde{u}\}.$$

This model is not quite a game theoretic one because the market makers do not explicitly maximize any particular objective. We could, however, replace the market efficiency condition in step two with an explicit Bertrand auction between at least two risk neutral bidders, each of whom observes the "order flow"  $\tilde{x} + \tilde{u}$  and nothing else. The result of this explicit auction procedure would be our market efficiency condition, in which profits of market makers are driven to zero. Modelling how market makers can earn the positive frictional profits necessary to attract them into the business of market making is an interesting topic which takes us away from our main objective of studying how price formation is influenced by the optimizing behavior of an insider in a somewhat idealized setting. Kyle [5], however, discusses a model of imperfect competition among market makers, in which many insiders with different information participate.

The insider exploits his monopoly power by taking into account the effect the quantity he chooses to trade in step one is expected to have on the price established in step two. In doing so, he takes the rule market makers use to set prices in step two as given. He is not allowed to influence this rule by committing to a particular strategy in step one: The quantity he trades is required to be optimal, given his information set at the time it is chosen. This requirement seems to be reasonable given anonymous trading and the strong incentives informed traders have to cheat given any other strategy they commit to. The insider is not allowed to condition the quantity he trades on price. A model in which insiders choose demand functions ("limit orders") instead of quantities ("market orders") is considered in Kyle [6].

Fortuitously, our model has an analytically tractable equilibrium in which the rules  $X$  and  $P$  are simple linear functions, as we show in the following theorem:

**THEOREM 1:** *There exists a unique equilibrium in which  $X$  and  $P$  are linear functions. Defining constants  $\beta$  and  $\lambda$  by  $\beta = (\sigma_u^2 / \Sigma_0)^{1/2}$  and  $\lambda = 2(\sigma_u^2 / \Sigma_0)^{-1/2}$ , the equilibrium  $P$  and  $X$  are given by*

$$(2.3) \quad X(\tilde{v}) = \beta(\tilde{v} - p_0), \quad P(\tilde{x} + \tilde{u}) = p_0 + \lambda(\tilde{x} + \tilde{u}).$$

*Proof:* Suppose that for constants  $\mu, \lambda, \alpha, \beta$ , linear functions  $P$  and  $X$  are given by

$$(2.4) \quad P(y) = \mu + \lambda y, \quad X(v) = \alpha + \beta v.$$

Given the linear rule  $P$ , profits can be written

$$(2.5) \quad E\{[\tilde{v} - P(x + \tilde{u})]x | \tilde{v} = v\} = (v - \mu - \lambda x)x.$$

Profit maximization of this quadratic objective requires that  $x$  solve  $v - \mu - 2\lambda x = 0$ . We thus have  $X(v) = \alpha + \beta v$  with

$$(2.6) \quad 1/\beta = 2\lambda, \quad \alpha = -\mu\beta.$$

Note that the quadratic objective (implied by the linear pricing rule  $P$ ) rules out mixed strategies and also makes linear strategies optimal even when nonlinear strategies are allowed.

Given linear  $X$  and  $P$ , the market efficiency condition is equivalent to

$$(2.7) \quad \mu + \lambda y = E\{\tilde{v} | \alpha + \beta\tilde{v} + \tilde{u} = y\}.$$

Normality makes the regression linear and application of the projection theorem yields

$$(2.8) \quad \lambda = \frac{\beta\Sigma_0}{\beta^2\Sigma_0 + \sigma_u^2}, \quad \mu - p_0 = -\lambda(\alpha + \beta p_0).$$

Solving (2.6) and (2.8) subject to the second order condition  $\lambda > 0$  yields the desired result. Note that we have  $\mu = p_0$ ,  $\alpha = -\beta p_0$ , and the second order condition rules out a solution with  $\beta$  and  $\lambda$  both negative. This completes the proof of the theorem.

**Properties of the Equilibrium.** The equilibrium  $X$  and  $P$  are determined by the exogenous parameters  $\Sigma_0$  and  $\sigma_u^2$ . To obtain a measure of the informativeness of prices, define  $\Sigma_1$  by  $\Sigma_1 = \text{var}\{\tilde{v} | \tilde{p}\}$ . A simple calculation shows that  $\Sigma_1 = \frac{1}{2}\Sigma_0$ ; thus, one-half of the insider's private information is incorporated into prices and the volatility of prices is unaffected by the level of noise trading  $\sigma_u^2$ .

The quantity  $1/\lambda$  measures the “depth” of the market, i.e. the order flow necessary to induce prices to rise or fall by one dollar. This measure of market liquidity is proportional to a ratio of the amount of noise trading to the amount of private information the informed trader is expected to have. In this sense, it

captures Bagehot's [1] intuition that market makers compensate themselves for bad trades due to the adverse selection of insiders by making the market less liquid. Maximized profits, given by  $v^2/(4\lambda)$ , are proportional to the depth of the market, because a proportional horizontal expansion of the supply curve induces the monopsonistic insider to trade a proportionately larger quantity without affecting prices, and this makes his profits correspondingly larger as well. Since an increase in noise trading brings forth more informed trading, it does not destabilize prices (a result which would disappear if the insider were risk averse). The expected profits of the insider (unconditional on  $\tilde{v}$ ) are given by  $E(\tilde{\pi}) = \frac{1}{2}(\Sigma_0\sigma_u^2)^{1/2}$ . The insider's profits are proportional to the standard deviations of both  $\tilde{v}$  and  $\tilde{u}$ .

As shown below, many of these properties generalize to the sequential auction model in an appropriate way.

### 3. A SEQUENTIAL AUCTION EQUILIBRIUM

In this section we generalize the model of one-shot trading by examining a model in which a number of auctions, or rounds of trading, take place sequentially. The resulting dynamic model is structured so that equilibrium prices at each auction reflect the information contained in the past and current order flow and so that the insider maximizes his expected profits, taking into account his effect on prices in both the current auction and in future auctions.

*Structure and Notation.* Trading takes place over one trading day, which begins at time  $t = 0$  and ends at time  $t = 1$ . There are  $N$  auctions, with  $t_n$  denoting the time at which the  $n$ th auction takes place. We assume

$$(3.1) \quad 0 = t_0 < t_1 < \cdots < t_N = 1,$$

so the sequence of auction dates  $\langle t_n \rangle$  partitions the interval  $[0, 1]$ .

Let  $\tilde{u}(t)$  denote a Brownian motion process with instantaneous variance  $\sigma_u^2$ , and define  $\tilde{u}_n$  and  $\Delta\tilde{u}_n$  by  $\tilde{u}_n = \tilde{u}(t_n)$  and  $\Delta\tilde{u}_n = \tilde{u}_n - \tilde{u}_{n-1}$ . We assume that the quantity traded by noise traders at the  $n$ th auction is  $\Delta\tilde{u}_n$ . The Brownian motion assumption implies that  $\Delta\tilde{u}_n$  is normally distributed with zero mean and variance  $\sigma_u^2 \Delta t_n$ , where  $\Delta t_n = t_n - t_{n-1}$ , and that the quantity traded at one auction is independent of the quantity traded at other auctions. The liquidation value of the asset,  $\tilde{v}$ , is still assumed to be normally distributed with mean  $p_0$  and variance  $\Sigma_0$ . The random variable  $\tilde{v}$  is distributed independently of the entire process  $\tilde{u}(t)$ .

The  $N$  auctions take place sequentially. Let  $\tilde{x}_n$  denote the aggregate position of the insider after the  $n$ th auction, so that  $\Delta\tilde{x}_n$  (defined by  $\Delta\tilde{x}_n = \tilde{x}_n - \tilde{x}_{n-1}$ ) denotes the quantity traded by the insider at the  $n$ th auction. Let  $\tilde{p}_n$  denote the market clearing price at the  $n$ th auction. At each auction, trade is structured in two steps as before, with information sets modified to take into account relevant information from past auctions. Since mixed trading strategies and random pricing rules are not optimal in what follows, we are justified in interpreting the trading rules and pricing rules as functions of the relevant observations. According to this interpretation, when the insider chooses the quantity to trade at step one of

an auction, he not only observes the liquidation value of the asset,  $\tilde{v}$ , but also past prices as well. Accordingly, for some measurable function  $X_n$ , his position after the  $n$ th auction is given by

$$(3.2) \quad \tilde{x}_n = X_n(\tilde{p}_1, \dots, \tilde{p}_{n-1}, \tilde{v}) \quad (n = 1, \dots, N),$$

from which the actual quantity traded is easily determined using information in the information set. When market makers set a market clearing price at step two of the  $n$ th auction, they not only observe the current value of the order flow,  $\Delta\tilde{x}_n + \Delta\tilde{u}_n$ , but they observe past values of the order flow as well. Accordingly, for some measurable function  $P_n$ , the price  $\tilde{p}_n$  is determined by

$$(3.3) \quad \tilde{p}_n = P_n(\tilde{x}_1 + \tilde{u}_1, \dots, \tilde{x}_n + \tilde{u}_n) \quad (n = 1, \dots, N).$$

Note that in the absence of mixed strategies, the insider can infer from his information set the quantities he has traded at past auctions, and the market makers can infer from their information set the prices they have set in the past. Note also that the insider can infer the quantities traded by noise traders in the past if the functions  $P_n$  are monotonic in their last arguments.

Now define the vectors of functions  $X$  and  $P$  by

$$(3.4) \quad X = \langle X_1, \dots, X_N \rangle, \quad P = \langle P_1, \dots, P_N \rangle.$$

We refer to  $X$  as the informed trader's "trading strategy" and to  $P$  as the market makers' "pricing rule."

For  $n = 1, \dots, N$ , let  $\tilde{\pi}_n$  denote the profits of the insider on positions acquired at auctions  $n, \dots, N$ . Clearly,  $\tilde{\pi}_n$  is given by

$$(3.5) \quad \tilde{\pi}_n = \sum_{k=n}^N (\tilde{v} - \tilde{p}_k) \tilde{x}_k \quad (n = 1, \dots, N).$$

To emphasize the dependence of  $\tilde{p}_n$ ,  $\tilde{x}_n$ , and  $\tilde{\pi}_n$  on  $P$  and  $X$ , we sometimes write

$$(3.6) \quad \tilde{p}_n = \tilde{p}_n(X, P), \quad \tilde{x}_n = \tilde{x}_n(X, P), \quad \tilde{\pi}_n = \tilde{\pi}_n(X, P).$$

*Equilibrium.* A *sequential auction equilibrium* is defined as a pair  $X, P$  such that the following two conditions hold:

1. *Profit Maximization:* For all  $n = 1, \dots, N$  and for all  $X' = \langle X'_1, \dots, X'_N \rangle$  such that  $X'_1 = X_1, \dots, X'_{n-1} = X_{n-1}$ , we have

$$(3.7) \quad E\{\tilde{\pi}_n(X, P) | \tilde{p}_1, \dots, \tilde{p}_{n-1}, \tilde{v}\} \geq E\{\tilde{\pi}_n(X', P) | \tilde{p}_1, \dots, \tilde{p}_{n-1}, \tilde{v}\}.$$

2. *Market Efficiency:* For all  $n = 1, \dots, N$ , we have

$$(3.8) \quad \tilde{p}_n = E\{\tilde{v} | \tilde{x}_1 + \tilde{u}_1, \dots, \tilde{x}_n + \tilde{u}_n\}.$$

A *linear equilibrium* is defined as a (sequential auction) equilibrium in which the component functions of  $X$  and  $P$  are linear, and a *recursive linear equilibrium* is defined as a linear equilibrium in which there exist constants  $\lambda_1, \dots, \lambda_N$  such that for  $n = 1, \dots, N$ ,

$$(3.9) \quad \tilde{p}_n = \tilde{p}_{n-1} + \lambda_n(\Delta\tilde{x}_n + \Delta\tilde{u}_n).$$

The market efficiency condition implies that trading prices follow a martingale whose pattern of volatility over time reflects the rate at which information is incorporated into prices. In a linear equilibrium, price increments are normally and independently distributed with zero means; thus, the distribution function for the pricing process is characterized by a sequence of variance parameters measuring the volatility of price fluctuations from auction to auction.

The profit maximization condition gives our equilibrium the flavor of a sequential equilibrium (as discussed by Kreps and Wilson [4]). The quantity  $\tilde{x}_n$  chosen at the  $n$ th auction maximizes expected profits over the remaining rounds of trading given the information available to the insider when he chooses it. There is no commitment to strategies. This means that the insider cannot influence the pricing rule by committing to a trading rule before prices are established. Conversely, while the market makers impute a trading strategy to the insider, they do not observe it; they only observe the order flow. Note, however, that the profit maximization condition implies that for all trading strategies  $X'$ ,

$$(3.10) \quad E\{\tilde{\pi}_n(X, P)\} \geq E\{\tilde{\pi}_n(X', P)\}.$$

*Characterization of Equilibrium.* In the rest of this section, we prove existence of a unique linear equilibrium, show that it is a recursive linear equilibrium, and characterize it as the solution to a difference equation system subject to boundary conditions. We suspect, but have not been able to prove, that equilibria with nonlinear  $X_n$  and  $P_n$  do not exist.

**THEOREM 2:** *There exists a unique linear equilibrium and this equilibrium is a recursive linear equilibrium. In this equilibrium there are constants  $\beta_n, \lambda_n, \alpha_n, \delta_n$ , and  $\Sigma_n$  such that for*

$$(3.11) \quad \Delta \tilde{x}_n = \beta_n (\tilde{v} - \tilde{p}_{n-1}) \Delta t_n,$$

$$(3.12) \quad \Delta \tilde{p}_n = \lambda_n (\Delta \tilde{x}_n + \Delta \tilde{u}_n),$$

$$(3.13) \quad \Sigma_n = \text{var}(\tilde{v} | \Delta \tilde{x}_1 + \Delta \tilde{u}_1, \dots, \Delta \tilde{x}_n + \Delta \tilde{u}_n),$$

$$(3.14) \quad E\{\tilde{\pi}_n | p_1, \dots, p_{n-1}, v\} = \alpha_{n-1}(v - p_{n-1})^2 + \delta_{n-1} \quad (n = 1, \dots, N).$$

*Given  $\Sigma_0$ , the constants  $\beta_n, \lambda_n, \alpha_n, \delta_n, \Sigma_n$  are the unique solution to the difference equation system*

$$(3.15) \quad \alpha_{n-1} = \frac{1}{4\lambda_n(1 - \alpha_n\lambda_n)},$$

$$(3.16) \quad \delta_{n-1} = \delta_n + \alpha_n \lambda_n^2 \sigma_u^2 \Delta t_n,$$

$$(3.17) \quad \beta_n \Delta t_n = \frac{1 - 2\alpha_n\lambda_n}{2\lambda_n(1 - \alpha_n\lambda_n)},$$

$$(3.18) \quad \lambda_n = \beta_n \Sigma_n / \sigma_u^2,$$

$$(3.19) \quad \Sigma_n = (1 - \beta_n \lambda_n \Delta t_n) \Sigma_{n-1} \quad (n = 1, \dots, N),$$

subject to  $\alpha_N = \delta_N = 0$  and the second order condition

$$(3.20) \quad \lambda_n(1 - \alpha_n \lambda_n) > 0.$$

REMARK: The parameters  $\beta_n (n = 1, \dots, N)$ , which characterize the insider's trading strategy  $X_n$ , measure the intensity with which the insider trades on the basis of his private observation, and the parameters  $\lambda_n (n = 1, \dots, N)$ , which characterize the recursive pricing rule, measure the depth of the market (with small  $\lambda_n$  corresponding to a deep market). The parameters  $\Sigma_n (n = 1, \dots, N)$ , which give the error variance of prices after the  $n$ th auction, measure how much of the insider's private information is not yet incorporated into prices (as estimated by market makers). Note that  $\Sigma_0$  is just the variance of the initial prior price  $p_0$ . The parameters  $\alpha_{n-1}$  and  $\delta_{n-1}$  define a quadratic profit function which gives the value of trading opportunities at auctions  $n, \dots, N$ .

*Outline of Proof:* The proof of the theorem is divided into three steps. In the first step, which is the most important one, a backward induction argument is used to obtain the insider's trading strategy and expected trading profits as a function of the pricing rule. Since the pricing rule is characterized by the market depth parameters  $\lambda_n$ , the insider's problem is intuitively one of deciding how intensely to trade on the basis of his private information, given the pattern of market depth expected at current and future auctions. If market depth at future auctions is greater than market depth at the current auction, the insider has an incentive to "save" his private information by trading small quantities now and large quantities later. Conversely, if market depth declines in future auctions, the insider has an incentive to trade intensely at the current auction, where profits are greater.

Intuitively, the second order condition (3.20) rules out a situation in which the insider can make unbounded profits by first destabilizing prices with unprofitable trades made at the  $n$ th auction, then recouping the losses and much more with profitable trades made at future auctions. When  $\lambda_n$  is large, it does not cost much to destabilize prices at the  $n$ th auction (because trading small quantities is sufficient), but when  $\alpha_n$  is large, the value of future trading opportunities to the insider from moving the price far away from its liquidation value is large. The second order condition accordingly rules out unbounded destabilization schemes by placing an upper bound on  $\lambda_n$  which decreases in  $\alpha_n$ .

The backward induction argument in step one of the proof simultaneously shows that the insider's profit function is quadratic and that the linear equilibrium is recursive. In addition, it shows explicitly how the parameter  $\alpha_n$ , which measures the value of private information at future auctions  $n + 1, \dots, N$  as a function of market depth at those auctions, combines with the current market depth parameter  $\lambda_n$  to generate via backward induction values of  $\beta_n$  and  $\alpha_{n-1}$ .

In step two of the proof, the market efficiency condition is used to derive  $\lambda_n$  and  $\Sigma_n$  from  $\beta_n$  and  $\Sigma_{n-1}$ . The idea here is that, given the level of noise trading ( $\sigma_u^2 \Delta t_n$ ), the depth of the market at a particular auction ( $\lambda_n$ ) depends negatively upon how much private information the insider has ( $\Sigma_{n-1}$ ) and how intensely

the insider trades upon the basis of his private information ( $\beta_n$ ), and this also determines how much of the insider's remaining private information is revealed at the particular auction and how much still remains private ( $\Sigma_n$ ). This step of the proof makes precise Bagehot's idea that market makers respond to insider trading by reducing the liquidity of the market.

In step three of the proof, it is shown that the relationships derived in the first two steps generate a difference equation system which characterizes the unique linear equilibrium.

**PROOF:** We now give the details of the three steps of the proof.

*Step 1.* To prove by backward induction that the informed trader's expected profits are of the quadratic form specified in (3.14), we begin with the boundary condition  $\alpha_N = \delta_N = 0$ , which states that no profits on new positions are made after trade is completed. Now make the inductive hypothesis that for constants  $\alpha_n$  and  $\delta_n$ , we have

$$(3.21) \quad E\{\tilde{\pi}_{n+1}(X, P) | p_1, \dots, p_n, v\} = \alpha_n (v - p_n)^2 + \delta_n.$$

Since  $\tilde{\pi}_n$  is given recursively by  $\tilde{\pi}_n = (\tilde{v} - \tilde{p}_n) \Delta \tilde{x}_n + \tilde{\pi}_{n+1}$ , we obtain

$$(3.22) \quad E\{\tilde{\pi}_n | p_1, \dots, p_{n-1}, v\} = \max_{\Delta x} E\{(\tilde{v} - \tilde{p}_n) \Delta x + \alpha_n (\tilde{v} - \tilde{p}_n)^2 + \delta_n | p_1, \dots, p_{n-1}, v\}.$$

In a linear equilibrium,  $p_n$  is given by

$$(3.23) \quad p_n = p_{n-1} + \lambda_n (\Delta x_n + \Delta u_n) + h,$$

where  $h$  is some linear function of  $\Delta x_1 + \Delta u_1, \dots, \Delta x_{n-1} + \Delta u_{n-1}$ . Plugging (3.23) into (3.22) and evaluating the conditional expectation yields

$$(3.24) \quad E\{\tilde{\pi}_n | p_1, \dots, p_{n-1}, v\} = \max_{\Delta x} \{(v - p_{n-1} - \lambda_n \Delta x - h) \Delta x + \alpha_n (v - p_{n-1} - \lambda_n \Delta x - h)^2 + \alpha_n \lambda_n^2 \sigma_u^2 \Delta t_n + \delta_n\}.$$

Since maximized profits are quadratic in  $\Delta x$ , the maximizing  $\Delta x$  (which we write  $\Delta x_n$ ) is easily shown to be given by

$$(3.25) \quad \Delta x_n = \beta_n (v - p_{n-1} - h) \Delta t_n,$$

where  $\beta_n \Delta t_n$  is given by (3.17). The second order condition is (3.20).

We now claim that  $\tilde{h} = 0$ . To prove this, observe that the market efficiency condition implies

$$(3.26) \quad E\{\Delta \tilde{p}_n | \Delta \tilde{x}_1 + \Delta \tilde{u}_1, \dots, \Delta \tilde{x}_{n-1} + \Delta \tilde{u}_{n-1}\} = 0.$$

An explicit calculation shows, however, that

$$(3.27) \quad E\{\Delta \tilde{p}_n | \Delta \tilde{x}_1 + \Delta \tilde{u}_1, \dots, \Delta \tilde{x}_{n-1} + \Delta \tilde{u}_{n-1}\} = \frac{\tilde{h}}{2(1 - \alpha_n \lambda_n)}.$$



Clearly, then, we must have  $\tilde{h}=0$  (with probability one). From this it follows from (3.23) and (3.25) that  $\Delta\tilde{p}_n$  and  $\Delta\tilde{x}_n$  have the recursive form given by (3.11) and (3.12), as stated in the theorem. Furthermore, it is also easy to show from (3.24) that  $\alpha_{n-1}$  and  $\delta_{n-1}$  are given by (3.15) and (3.16).

*Step 2.* Consider the values of  $\lambda_n$  and  $\Sigma_n$  consistent with the market efficiency condition. Because  $\tilde{v}-\tilde{p}_{n-1}$  is independent from  $\Delta\tilde{x}_1+\Delta\tilde{u}_1, \dots, \Delta\tilde{x}_{n-1}+\Delta\tilde{u}_{n-1}$ , we obtain from the market efficiency condition

$$(3.28) \quad \tilde{p}_n - \tilde{p}_{n-1} = E\{\tilde{v} - \tilde{p}_{n-1} | \Delta\tilde{x}_n + \Delta\tilde{u}_n\}.$$

Using (3.11), a simple application of the projection theorem for normally distributed random variables confirms that  $\Delta\tilde{p}_n$  is indeed of the form specified in (3.12), as was shown necessary in step one, and yields the following explicit expressions for  $\lambda_n$  and  $\Sigma_n$ :

$$(3.29) \quad \lambda_n = \frac{\beta_n \Sigma_{n-1}}{\beta_n^2 \Sigma_{n-1} \Delta t_n + \sigma_u^2},$$

$$\Sigma_n = \frac{\sigma_u^2 \Sigma_{n-1}}{\beta_n^2 \Sigma_{n-1} \Delta t_n + \sigma_u^2}.$$

These are equivalent to (3.18) and (3.19).

*Step 3.* In steps one and two above, it is shown that given  $\Sigma_0$ , equations (3.11)–(3.20) and the boundary condition  $\alpha_N = \delta_N = 0$  are necessary for a linear equilibrium. Clearly, they are also sufficient. We now show that given  $\Sigma_0$ , the difference equation system (3.15)–(3.19) has a unique solution satisfying the boundary condition  $\alpha_N = \delta_N = 0$  and the second order condition (3.20); it thus characterizes the unique linear equilibrium.

We first claim that given nonnegative values of  $\alpha_n$  and  $\Sigma_n$ , there is a unique way to iterate the system backwards such that the second order condition is satisfied. To prove this, combine equations (3.18) with (3.17) and simplify, obtaining

$$(3.30) \quad (1 - \lambda_n^2 \sigma_u^2 \Delta t_n / \Sigma_n)(1 - \alpha_n \lambda_n) = \frac{1}{2}.$$

Given nonnegative  $\alpha_n$  and  $\Sigma_n$ , this is a cubic equation in  $\lambda_n$ , which has three real roots. While neither the largest nor the smallest root satisfies the second order condition, the middle root does satisfy the second order condition. Thus,  $\lambda_n$  is uniquely defined by (3.30). Furthermore, given  $\lambda_n$ , it is a simple matter to obtain  $\beta_n$ ,  $\alpha_{n-1}$ ,  $\delta_{n-1}$ , and  $\Sigma_{n-1}$  from (3.15)–(3.19), and we have therefore iterated the system backwards one step.

Given the boundary condition  $\alpha_N = \delta_N = 0$ , the backward iteration procedure defines a family of solutions to the difference equation system parametrized by the terminal value  $\Sigma_N$  used to start off the backward iteration. We now claim that only one terminal value exists such that the correct initial value  $\Sigma_0$  is obtained at the last step of the backward iteration. To prove this, observe that if  $\alpha_{n-1}$ ,  $\delta_{n-1}$ ,  $\Sigma_{n-1}$ ,  $\beta_n$ ,  $\lambda_n$  ( $n = 1, \dots, N$ ) solve the difference equation system given arbitrary  $\Sigma_N$  and  $\alpha_N = \delta_N = 0$ , then for any positive constant  $\zeta$ , it is also true that

$\zeta\alpha_{n-1}, \zeta\delta_{n-1}, \zeta^2\Sigma_{n-1}, \zeta\beta_n, \zeta^{-1}\lambda_n$  ( $n = 1, \dots, N$ ) solve the difference equation system when the terminal value is  $\zeta^2\Sigma_N$ . Since this implies that  $\Sigma_N$  is proportional to  $\Sigma_0$  in any solution, there is a unique solution for any initial value  $\Sigma_0$ . This completes the proof of the theorem.

*Propriétés of the Equilibrium.* It is apparent from inspecting the difference equation system that many of the properties of the single auction equilibrium generalize to the sequential auction equilibrium. From (3.19), it is clear that the parameter  $\Sigma_n$ , which measures the informativeness of prices, declines monotonically, reflecting the fact that information is gradually incorporated into prices. While we have  $\Sigma_N > 0$  (so not all information is incorporated into prices by the end of trading), we show below that  $\Sigma_N$  may be very small. It is clear from inspection of the difference equation system that if  $\sigma_u$  doubles, then  $\lambda_n$  halves;  $\alpha_n, \delta_n$ , and  $\beta_n$  double; and  $\Sigma_n$  is unaffected. Thus, increasing the amount of noise trading increases market depth proportionately, increases proportionately the profits of the insider by encouraging him to trade more, and leaves the informativeness of prices unchanged. It is clear from the proof that if the amount of prior inside information, as measured by  $\Sigma_0^{1/2}$ , increases, then market depth decreases proportionately but expected (ex ante) profits increase proportionately. Thus, profits are proportional to  $(\Sigma_0\sigma_u^2)^{1/2}$ , as in the single auction model.

The rest of this paper investigates the properties of the equilibrium in more detail by examining what happens to the equilibrium when the interval between auctions becomes very small. In particular, we are interested in learning more about the dynamic behavior of  $\Sigma_n$  and  $\lambda_n$ .

## 5. A CONTINUOUS AUCTION EQUILIBRIUM

In this section we discuss unrigorously a model in which trading takes place continuously rather than at discrete intervals. Our main results are that the depth of the market is constant over time and the volatility of prices is constant. In Section 6, we show rigorously that the sequential auction equilibrium converges to the continuous auction equilibrium discussed here when auctions are held frequently.

Proceeding intuitively, let us define a *continuous auction equilibrium* exactly analogously to the sequential auction equilibrium discussed above. We take it for granted that a unique linear equilibrium with a structure analogous to the recursive equilibrium Theorem 2 exists. Write the analogues of (3.5), (3.11), and (3.12) as

$$(4.1) \quad d\pi(t) = [v - p(t) - dp(t)] dx(t) = [v - p(t)] dx(t),$$

$$(4.2) \quad dx(t) = \beta(t)[v - p(t)] dt,$$

$$(4.3) \quad dp(t) = \lambda(t)[dx(t) + du(t)].$$

Note that in this notation, we suppress tildes over random variables. Here we assume that the trading strategy  $dx$  and the pricing rule  $dp$  are characterized by

the functions  $\beta(\cdot)$  and  $\lambda(\cdot)$ , respectively; we will not worry about trading strategies which have a more complicated structure. Readers unwilling to *assume* that a linear equilibrium has the simple structure of (4.1)–(4.3) are invited to interpret the following discussion as referring to a modified equilibrium concept in which  $\pi$ ,  $x$ , and  $p$  are *constrained* to have the simple linear form which we give them here.

Because the  $dp\,dx$  term in (4.1) is of order  $dt^{3/2}$ , whether trades are priced at the beginning or the end of the instant in which they occur has an inconsequential effect on the profits of the insider, given that he buys and sells smoothly. If trades were priced at the beginning of the instant in which they occurred, however, the insider would not want to trade smoothly: Instead, he would want to trade unbounded quantities at the prices quoted by market makers in advance, since his quantity traded would have no effect on the immediate execution price. Thus, we make the assumption that trades are priced at the end of the instant in which they occur, which is consistent with the sequential auction equilibrium discussed above. The assumption as to whether pricing occurs at the beginning or the end of the instant in which trades occur does affect the profits of the market makers and the noise traders because we have  $du\,dp = -\lambda\sigma_u^2\,dt$ , i.e.,  $du\,dp$  is of order  $dt$  and not of higher order like the corresponding term for the insider. “End-of-instant pricing” makes the market efficiency condition a zero profit condition for market makers by requiring that noise traders bear the losses incurred by virtue of the fact that they drive prices against themselves as they trade.

In a continuous auction equilibrium, the profit maximization condition states that given a pricing rule, the trading rule  $dx$  maximizes expected profits given by

$$(4.4) \quad E\{\pi(t) | \langle p(s) \rangle_{s \in [0, t]}, v\} = E\left\{ \int_{s=t}^1 d\pi(s) \mid \langle p(s) \rangle_{s \in [0, t]}, v \right\}.$$

The market efficiency condition states that the pricing rule satisfies

$$(4.5) \quad E\{v | \langle dx + du \rangle_{s \in [0, t]}\} = p(t).$$

Given arbitrary  $\beta(\cdot)$  and  $\lambda(\cdot)$ , define  $\Sigma^*(t)$  and  $\Sigma(t)$  by

$$(4.6) \quad \Sigma^*(t) = E\{[v - p(t)]^2\},$$

$$(4.7) \quad \Sigma(t) = \text{var}\{\tilde{v} | \langle dx + du \rangle_{s \in [0, t]}\}.$$

Clearly, the market efficiency condition implies  $\Sigma^*(t) = \Sigma(t)$ .

Assuming that the equilibrium has the linear form given in (3.2) and (3.3), we obtain the following theorem:

**THEOREM 3:** *In the recursive continuous auction equilibrium, the function  $\lambda(t)$  is a constant given by*

$$(4.8) \quad \lambda(t) = (\Sigma_0 / \sigma_u^2)^{1/2},$$

and the functions  $\Sigma(t)$ ,  $\beta(t)$ ,  $\alpha(t)$ , and  $\delta(t)$  are given by

$$(4.9) \quad \Sigma(t) = (1-t)\Sigma_0,$$

$$(4.10) \quad \beta(t) = \sigma_u^2 \lambda(t) / \Sigma(t) = \sigma_u \Sigma_0^{-1/2} / (1-t),$$

$$(4.11) \quad \alpha(t) = \frac{1}{2}(\sigma_u^2 / \Sigma_0)^{1/2}, \quad t \in (0, 1),$$

$$(4.12) \quad \delta(t) = \frac{1}{2}(\sigma_u^2 \Sigma_0)^{1/2}(1-t).$$

PROOF: Let us first examine the optimal trading rule given an arbitrary pricing rule characterized by some function  $\lambda(t)$ . Since the trading rule is assumed to be linear, we need only maximize ex ante profits, given by

$$(4.13) \quad E\{\pi(0)\} = E\left\{\int_{t=0}^1 d\pi(t)\right\}.$$

From (4.1)–(4.3), we obtain

$$(4.14) \quad E\{d\pi(t)\} = \beta(t)\Sigma^*(t) dt.$$

We also have

$$(4.15) \quad \Sigma^*(t+dt) = E\{(v-p-dp)^2\} = (1-\lambda\beta dt)^2\Sigma^*(t) + \lambda^2\sigma_u^2 dt,$$

from which we obtain

$$(4.16) \quad d\Sigma^*/dt = -2\lambda\beta\Sigma^*(t) + \lambda^2\sigma_u^2.$$

Plugging (4.14) into (4.13), we obtain

$$(4.17) \quad E\{\pi(0)\} = \int_{t=0}^1 \beta(t)\Sigma^*(t) dt,$$

where  $\Sigma^*(t)$  evolves according to the differential equation (4.16). Now (4.16) is equivalent to

$$(4.18) \quad \beta(t)\Sigma^*(t) = \frac{1}{2}[\lambda(t)\sigma_u^2 - \Sigma^{*'}(t)/\lambda(t)],$$

and plugging this into (4.17) yields

$$(4.19) \quad E\{\pi(0)\} = \frac{1}{2} \int_{t=0}^1 \lambda(t)\sigma_u^2 dt + \frac{1}{2} \int_{t=0}^1 \lambda^{-1}(t) d(-\Sigma^*(t)).$$

In this equation, the control  $\beta(t)$  has been eliminated from the optimization problem and only the state  $\Sigma^*(t)$  remains. While  $\Sigma^*(t)$  is defined for all functions of bounded variation, it is clear that all such functions attainable (even in a limiting sense) with controls  $\beta(t)$  must satisfy  $\Sigma^*(t) \geq 0$ . Thus, the insider's problem is equivalent to choosing an (otherwise attainable)  $\Sigma^*(t)$  which satisfies this nonnegativity constraint.

We now show using (4.19) that only a constant function  $\lambda(t)$  is consistent with equilibrium. First, observe that we must have  $\Sigma(1) = 0$  if the right-hand-side of (4.19) is to be maximized; in other words, the price  $p(t)$  must be driven by the

insider to its underlying value  $v$  by the end of trading. Next, observe that if  $\lambda(t)$  ever decreases (i.e., if market depth ever increases), then unbounded profits can be generated by letting  $\Sigma(t)$  increase a large amount before the increase in  $\lambda(t)$  and then letting it decrease the same amount after the decrease. Since unbounded profits are inconsistent with equilibrium, we conclude that  $\lambda(t)$  must be monotonically nondecreasing in any equilibrium.

Intuitively, the requirement that  $\lambda(t)$  is nondecreasing eliminates profitable destabilization schemes and thus generalizes the second order condition (3.20) of the discrete model. With continuous trading, an insider who destabilizes prices by acquiring a large position in many small parcels over a short time period acts much like a perfectly discriminating monopsonist who moves up along a given supply curve. Since the supply curve is linear, the average price paid is approximately the mean of the highest and lowest prices paid on the small parcels. If the supply curve subsequently flattens (i.e., market depth increases), the insider can liquidate his position at a more favorable average price and thus generate unbounded profits by acquiring a large enough position in the first place.

It is also clear from inspecting (4.19) that in order to maximize profits, we must have  $\Sigma^*(t) = 0$  at a point where  $\lambda(t)$  is minimized. If  $\lambda(t)$  were ever to increase, we would therefore have  $\lambda(t^*) = 0$  for some  $t^*$  satisfying  $t^* < 0$ . From the market efficiency condition, this would imply that all information would be incorporated into prices before the end of trading and thus that prices would cease to fluctuate; but the only way for this to happen is to have  $\lambda(t) = 0$  for  $t^* < t < 1$  and this is inconsistent with  $\lambda(t)$  never decreasing. We conclude that  $\lambda(t)$  never increases either. We have thus proved that  $\lambda(t)$  must be a constant in equilibrium.

From (4.19), it is clear that if  $\lambda(t)$  is constant, any function  $\Sigma^*(t)$  satisfying  $\Sigma^*(1) = 0$  satisfies the profit maximization condition as long as it is attainable with some function  $\beta(t)$ . To calculate the values of  $\lambda$ ,  $\beta(t)$ , and  $\Sigma(t)$  consistent with equilibrium, we therefore turn to the market efficiency condition. Observe that (if  $\beta(t)$  is finite) the instantaneous variance of  $dp$  is  $\lambda^2 \sigma_u^2 dt$ , i.e., the volatility of prices is completely dominated by noise trading. In order to have  $\Sigma(1) = 0$  and market efficiency, the integral of volatilities must add up to the prior variance  $\Sigma_0$ . This gives us  $\lambda^2 \sigma_u^2 = \Sigma_0$ , from which (4.10) and (4.11) are immediate implications. To determine the values of  $\beta(t)$  consistent with market efficiency, observe that  $\beta(t)$  must be such that  $\lambda$  is the correct regression coefficient in the equation

$$(4.18) \quad E\{v - p(t) | \beta(t)[v - p(t)] dt + du(t)\} = \lambda[\beta(v - p(t)) dt + du(t)].$$

The appropriate Kalman filtering formula for the regression coefficient is  $\lambda = \beta(t)/\Sigma(t)$ , from which (4.10) is an immediate consequence. We leave (4.11) and (4.12) for the reader to derive. This completes the proof of the theorem.

*Properties of the Continuous Auction Equilibrium.* The fact that  $\Sigma'(t)$  is a constant (or, equivalently, that  $\lambda(t)$  is a constant) in a continuous auction equilibrium implies that trading prices have constant volatility over time and therefore that information is gradually incorporated into prices at a constant rate. From the

fact that  $\Sigma(1) = 0$ , we infer that all of the insider's private information is incorporated into prices by the end of trading, i.e.  $p(t)$  converges to  $v$  (in mean square) at  $t \rightarrow 1$ . Because of normality and the martingale properties inherent in the market efficiency condition, the price actually follows a Brownian motion process with instantaneous variance  $\Sigma_0$ . Of course, the insider knows that the price path will eventually converge to the liquidation value  $\tilde{v}$ , but to market makers, who do not observe  $\tilde{v}$  explicitly, price fluctuations appear to have no drift.

Note that since the volatility of prices is determined by noise traders and not by the insider, there is a sense in which the "trading volume" of the insider is small. Despite his small trading volume, however, the insider ultimately determines what price is established at the end of trading. He does this because his trades, unlike the trades of noise traders, are positively correlated from period to period.

The expected (ex ante) profits of the insider, which equal the expected losses of noise traders, can be shown to equal  $\Sigma_0^{1/2}\sigma_u$ . This is exactly double the profits the insider expects in the single auction equilibrium.

*Market Liquidity.* It is interesting to compare the liquidity properties of the continuous auction equilibrium with the corresponding properties of a sequential auction equilibrium. It was pointed out above that "market liquidity" refers to several different elements of transactions costs, including "tightness," "depth," and "resiliency."

"Tightness" refers to the cost of turning over a position in a short period of time. In the continuous auction equilibrium, the market is infinitely tight, in the sense that it is costless to turn over a position very quickly. This occurs because a trader acts like a perfectly discriminating monopsonist, who moves along a given "expected residual supply curve." In a sequential auction equilibrium, however, the monopolist is not able to trade at every price along the supply curve because auctions are not held closely enough together. As a result, the market is not infinitely tight, and the cost of turning over a position is an increasing function of how quickly it must be done.

"Depth" refers to the ability of the market to absorb quantities without having a large effect on price. It is measured by the reciprocal of the liquidity parameter  $\lambda(t)$ . In the continuous auction equilibrium, the depth of the market is constant. In the proof of Theorem 3, we showed that this is the case because neither increasing nor decreasing depth is consistent with behavior by the informed trader which is "stable" enough to sustain an equilibrium. If depth ever increases, the insider wants to destabilize prices (before the increase in depth) to generate unbounded profits. If depth ever decreases, the insider wants to incorporate all of his private information into the price immediately. This constancy of market depth also explains why the volatility of prices is constant in a continuous auction equilibrium. In a sequential auction equilibrium, depth is not constant over time.

Market "resiliency" refers to the speed with which prices tend to converge towards the underlying liquidation value of the commodity. Resiliency also

measures the rate at which prices bounce back from an uninformative shock. In both the continuous and sequential auction equilibria, the resiliency of prices is determined by the trading of the insider. Noise trading causes the price to wander aimlessly, with no tendency to return to an underlying value. Note that in the continuous auction equilibrium, the fact that  $\beta(t)$  increase in  $t$  means that the resiliency of the market increases in  $t$ . Since  $\beta(t) \rightarrow \infty$  as  $t \rightarrow 1$ , the market becomes infinitely resilient near the end of trading.

It is clear from this description of the liquidity characteristics of the market that a continuous auction equilibrium has essentially the same features Black uses to characterize a liquid market. What makes our model different from what is already in the literature is that these characteristics of market liquidity are results derived within an appropriate model of maximizing behavior.

## 5. A CONVERGENCE RESULT

How are the properties of the discrete model of sequential trading related to the properties of the continuous auction equilibrium when the interval between auctions is small? In this section, we answer this question by proving a convergence result. We show that as the interval between auctions in the discrete model becomes uniformly small, the unique sequential auction equilibrium characterized in Theorem 2 converges to the continuous auction equilibrium of Theorem 3.

Define the *mesh* of a partition of the interval  $[0, 1]$  into auction dates, which we denote  $|\Delta t|$ , as the maximum interval between auctions. Let  $\lambda_n, \beta_n, \Sigma_n, \alpha_n, \delta_n$  be defined as continuous functions  $\lambda(t), \beta(t)$ , etc., by the conventions  $\lambda(t) = \lambda_{n-1}$  for all  $t \in [t_{n-1}, t_n)$ , etc. We have the following theorem:

**THEOREM 4:** *Holding  $\Sigma_0$  and  $\sigma_u^2$  constant, consider a sequence of sequential auction equilibria (with different partitions defining auction dates) such that  $|\Delta t| \rightarrow 0$ . Then the values of  $\lambda(t), \beta(t), \Sigma(t), \alpha(t), \delta(t)$  characterized in Theorem 2 converge to the corresponding values in the continuous auction equilibrium obtained in Theorem 3. For  $\Sigma(t)$  and  $\delta(t)$ , the convergence is uniform for all  $t \in [0, 1]$ . For  $\lambda(t), \beta(t)$ , and  $\alpha(t)$  the convergence is uniform in all closed intervals which do not contain  $t = 1$ .*

**PROOF:** To prove this theorem, we would like to show that the difference equation system in Theorem 2 converges to a differential equation system. This approach does not lead to a simple proof, however, because the difference equation system is so badly behaved around  $t = 1$  that standard convergence theorems cannot be applied: Note, for example the discontinuity in  $\alpha(t)$  when  $t = 1$ , apparent from Theorem 3 and the boundary condition  $\alpha_N = 0$ . As an alternative, we show that the difference equation system in Theorem 2 can be tackled by first obtaining a difference equation in one variable (denoted  $\phi$  below), then characterizing explicitly the behavior of this difference equation, and finally using the limiting solution to this difference equation to determine the limiting behavior of the original difference equation system.



We can write equations (3.17) and (3.30), respectively as

$$(5.1) \quad \beta_n \lambda_n \Delta t_n = z_1 / (1 + z_1), \quad z_1 = 1 - 2\alpha_n \lambda_n,$$

$$(5.2) \quad \beta_n \lambda_n \Delta t_n = z_2 / (1 + z_2), \quad z_2 = \beta_n^2 \Sigma_{n-1} \Delta t_n / \sigma_u^2.$$

Equating  $z_1$  and  $z_2$  yields

$$(5.3) \quad 1 - 2\alpha_n \lambda_n = \beta_n^2 \Sigma_{n-1} \Delta t_n / \sigma_u^2.$$

Equations (3.17) and (3.15) are equivalent to

$$(5.4) \quad \beta_n \Delta t_n = 2\alpha_{n-1} (1 - 2\alpha_n \lambda_n),$$

$$(5.5) \quad (\alpha_n - \alpha_{n-1}) / \alpha_{n-1} = -(1 - 2\alpha_n \lambda_n)^2.$$

Now define  $\phi_n$  by

$$(5.6) \quad \phi_n = 4\alpha_n^2 \Sigma_n / \sigma_u^2 \quad (n = 0, \dots, N).$$

Combining (5.3) and (5.4) to eliminate  $\beta_n$  yields

$$(5.7) \quad 1 - 2\alpha_n \lambda_n = \Delta t_n / \phi_{n-1}.$$

Combining (5.7) with (5.3) and substituting into (3.31) yields

$$(5.8) \quad \Sigma_n / \Sigma_{n-1} = (1 + \Delta t_n / \phi_{n-1})^{-1},$$

and combining (5.5) with (5.7) yields

$$(5.9) \quad \alpha_n / \alpha_{n-1} = 1 - \Delta t_n^2 / \phi_{n-1}^2.$$

These two difference equations define  $\Sigma_n$  and  $\alpha_n$  in terms of  $\phi_{n-1}$ . Multiplying (5.8) and (5.9) together yields the following difference equation for  $\phi_n$  in terms of itself:

$$(5.10) \quad \phi_n / \phi_{n-1} = (1 - \Delta t_n^2 / \phi_{n-1}^2) [1 - \Delta t_n / (\phi_{n-1} + \Delta t_n)].$$

This can be simplified to

$$(5.11) \quad \phi_n - \phi_{n-1} = -\Delta t_n - \Delta t_n^2 / \phi_{n-1} + \Delta t_n^3 / \phi_{n-1}^2.$$

The boundary condition is  $\phi_N = 0$ . To iterate this difference equation for  $\phi_n$  backwards, a cubic equation must be solved at each step. We leave it to the reader to show that of the three solutions to this cubic equation, only one solution makes economic sense, and this solution satisfies

$$(5.12) \quad -5/4 < (\phi_n - \phi_{n-1}) / \Delta t_n < -1,$$

$$(5.13) \quad (\phi_n - \phi_{n-1}) / \Delta t_n \rightarrow -1 \quad \text{as} \quad \phi_n / \Delta t_n \rightarrow \infty.$$

From (5.12) and (5.13) it is clear that for the continuous version of  $\phi$ , denoted  $\phi(t)$ , we have

$$(5.14) \quad \phi(t) \rightarrow 1 - t,$$

and the convergence is uniform for all  $t \in [0, 1]$ . We have thus calculated the limiting behavior of  $\phi(t)$ .

To calculate the limiting behavior of  $\Sigma(t)$ , observe that we can use (5.8) to write

$$(5.15) \quad (\Sigma_n - \Sigma_{n-1})/\Sigma_{n-1} = -\Delta t_n/(1 - t_n) + o(|\Delta t|),$$

and standard convergence results for converting difference equations into differential equations allow us to conclude that the solution to (5.15) converges to the solution of the difference equation

$$(5.16) \quad \Sigma'/\Sigma = -1/(1 - t)$$

uniformly for all  $t$  bounded away from  $t = 1$ . Furthermore, the solution to (5.16) is

$$(5.17) \quad \Sigma(t) = (1 - t)\Sigma_0.$$

To calculate the limiting behavior of  $\alpha(t)$ , it is clear from the definition of  $\phi$  in (5.6) and from the limit results for  $\phi(t)$  and  $\Sigma(t)$  in (5.14) and (5.17) that for all  $t$  bounded away from  $t = 1$ ,  $\alpha(t)$  converges uniformly to the constant  $\frac{1}{2}\sigma_u/\Sigma_0^{1/2}$ . Convergence results for  $\beta$ ,  $\lambda$ , and  $\delta$  are simple exercises which we leave to the reader. This completes the proof of the theorem.

*Specific Examples.* In the process of proving the convergence result in Theorem 4, we also demonstrate a simple procedure for calculating solutions to the difference equation system which characterizes the sequential auction equilibrium. That procedure is to calculate  $\gamma_n$  from (5.10) by backward iteration from  $\gamma_N = 0$ , next to calculate  $\Sigma_n$  from (5.8) by forward iteration given  $\Sigma_0$ , then to use (5.6) to calculate  $\alpha_n$ , and finally to calculate other parameters from  $\alpha_n$  and  $\Sigma_n$ . In Figure 1, values of the liquidity parameter  $\lambda_n$  and error variance  $\Sigma_n$  are plotted for the particular cases  $N = 4$ ,  $N = 20$ , and  $N = 100$ . In these cases, exogenous parameters are normalized by setting  $\Sigma_0 = \sigma_u^2 = 1$ , and auctions are assumed to occur at equally spaced intervals. For the purpose of comparison, results for the continuous case ( $\Sigma_n = 1 - t$ ,  $\lambda_n = 1$ ) are also given. Figure 1 illustrates clearly convergence of parameters to the continuous model as  $N$  becomes large.

## 6. CONCLUSION

We have investigated a model of speculative trading in which an insider maximizes profits by exploiting strategically his monopoly power in a dynamic context. The model is important for a number of reasons.

It illustrates that modeling price innovations as functions of quantities traded is not inconsistent with modelling price innovations as the consequence of new information. Simultaneously, it illustrates that the strategic exercise of monopoly power by an insider is in no way inconsistent with prices being set efficiently in the semi-strong sense. The model shows how a discrete model of sequential trading (structured to resemble a sequential equilibrium) converges in the limit as trading takes place very frequently to a simple model of continuous trading.

1334

ALBERT S. KYLE

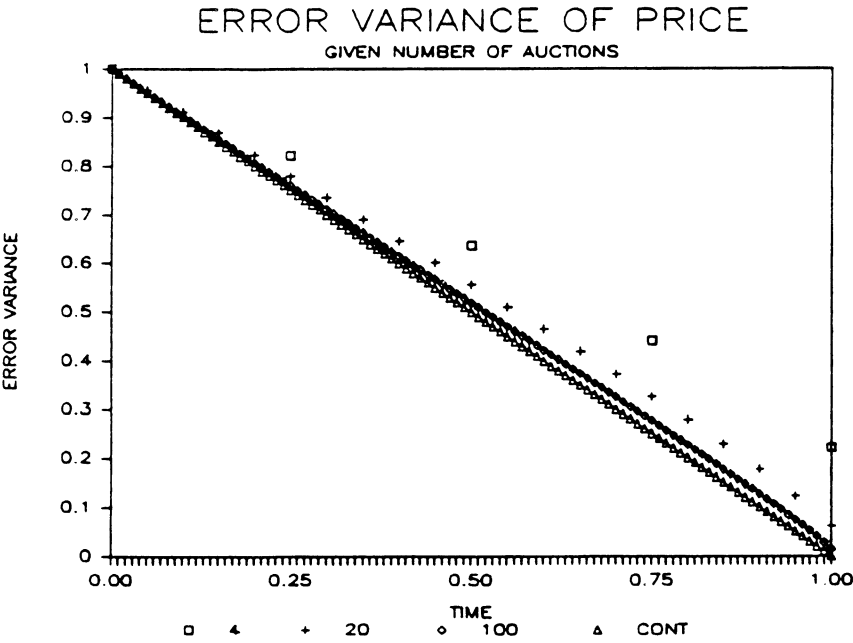
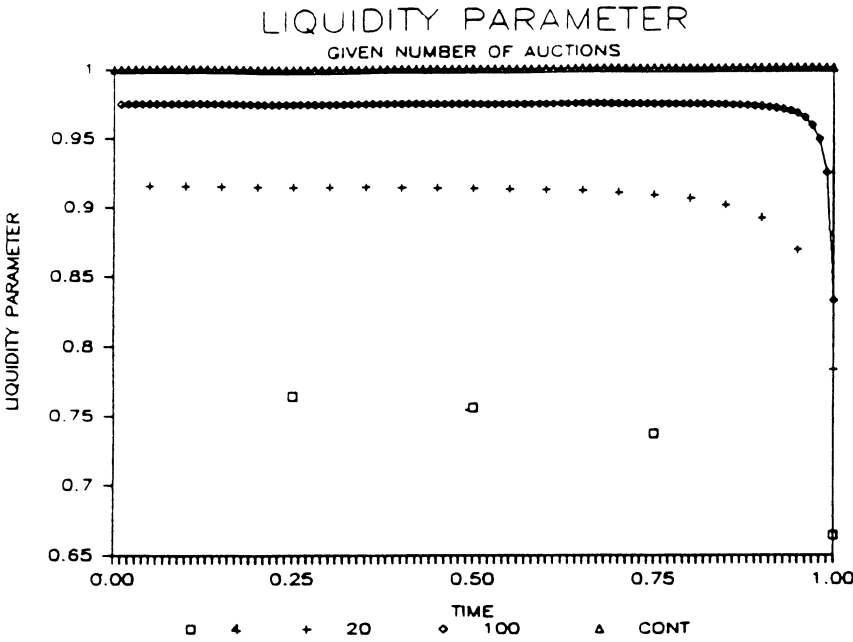


FIGURE 1

In doing so, it illustrates that constant volatility of prices need not require that the information upon which trades are based be produced in a smooth manner. Finally, the model demonstrates how the liquidity characteristics of an “efficient,” “frictionless” market can be derived from underlying information asymmetries in a dynamic trading environment which captures some relevant features of trading in organized exchanges.

*Princeton University*

*Manuscript received January 1984; revision received December, 1984.*

#### REFERENCES

- [1] BAGEHOT, WALTER: “The Only Game in Town,” *Financial Analysts Journal*, 27(1971), 12–14, 22.
- [2] BLACK, FISCHER: “Towards a Fully Automated Exchange, Part I,” *Financial Analysts Journal*, 27(1971), 29–34.
- [3] GLOSTEN, LAWRENCE R., AND PAUL R. MILGROM: “Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders,” J. L. Kellogg Graduate School of Management Working Paper No. 570, 1984.
- [4] KREPS, DAVID M., AND ROBERT WILSON: “Sequential Equilibria,” *Econometrica*, 50(1982), 863–894.
- [5] KYLE, ALBERT S.: “Market Structure, Information, Futures Markets, and Price Formation,” in *International Agricultural Trade: Advanced Readings in Price Formation, Market Structure, and Price Instability*, ed. by Gary G. Storey, Andrew Schmitz, and Alexander H. Sarris. Boulder and London: Westview Press, 1984, 45–64.
- [6] ———: “Equilibrium in a Speculative Market with Strategic Informed Trading,” unpublished manuscript, 1984.



# Exhibit 58

*Journal of Economic Literature*  
Vol. XXXV (March 1997), pp. 13-39

# Event Studies in Economics and Finance

A. CRAIG MACKINLAY

*The Wharton School, University of Pennsylvania*

*Thanks to John Campbell, Bruce Grundy, Andrew Lo, and two anonymous referees for helpful comments and discussion. Research support from the Rodney L. White Center for Financial Research is gratefully acknowledged.*

## 1. Introduction

ECONOMISTS are frequently asked to measure the effects of an economic event on the value of firms. On the surface this seems like a difficult task, but a measure can be constructed easily using an event study. Using financial market data, an event study measures the impact of a specific event on the value of a firm. The usefulness of such a study comes from the fact that, given rationality in the marketplace, the effects of an event will be reflected immediately in security prices. Thus a measure of the event's economic impact can be constructed using security prices observed over a relatively short time period. In contrast, direct productivity related measures may require many months or even years of observation.

The event study has many applications. In accounting and finance research, event studies have been applied to a variety of firm specific and economy wide events. Some examples include mergers and acquisitions, earnings announcements, issues of new debt or equity, and announcements of macroeconomic variables such as the trade

deficit.<sup>1</sup> However, applications in other fields are also abundant. For example, event studies are used in the field of law and economics to measure the impact on the value of a firm of a change in the regulatory environment (see G. William Schwert 1981) and in legal liability cases event studies are used to assess damages (see Mark Mitchell and Jeffry Netter 1994). In the majority of applications, the focus is the effect of an event on the price of a particular class of securities of the firm, most often common equity. In this paper the methodology is discussed in terms of applications that use common equity. However, event studies can be applied using debt securities with little modification.

Event studies have a long history. Perhaps the first published study is James Dolley (1933). In this work, he examines the price effects of stock splits, studying nominal price changes at the time of the split. Using a sample of 95 splits from 1921 to 1931, he finds that the price in-

<sup>1</sup> The first three examples will be discussed later in the paper. Grant McQueen and Vance Roley (1993) provide an illustration of the fourth using macroeconomic news announcements.



creased in 57 of the cases and the price declined in only 26 instances. Over the decades from the early 1930s until the late 1960s the level of sophistication of event studies increased. John H. Myers and Archie Bakay (1948), C. Austin Barker (1956, 1957, 1958), and John Ashley (1962) are examples of studies during this time period. The improvements included removing general stock market price movements and separating out confounding events. In the late 1960s seminal studies by Ray Ball and Philip Brown (1968) and Eugene Fama et al. (1969) introduced the methodology that is essentially the same as that which is in use today. Ball and Brown considered the information content of earnings, and Fama et al. studied the effects of stock splits after removing the effects of simultaneous dividend increases.

In the years since these pioneering studies, a number of modifications have been developed. These modifications relate to complications arising from violations of the statistical assumptions used in the early work and relate to adjustments in the design to accommodate more specific hypotheses. Useful papers which deal with the practical importance of many of the complications and adjustments are the work by Stephen Brown and Jerold Warner published in 1980 and 1985. The 1980 paper considers implementation issues for data sampled at a monthly interval and the 1985 paper deals with issues for daily data.

In this paper, event study methods are reviewed and summarized. The paper begins with discussion of one possible procedure for conducting an event study in Section 2. Section 3 sets up a sample event study which will be used to illustrate the methodology. Central to an event study is the measurement of an abnormal stock return. Section 4 details the first step—measuring the normal performance—and Section 5 follows

with the necessary tools for calculating an abnormal return, making statistical inferences about these returns, and aggregating over many event observations. The null hypothesis that the event has no impact on the distribution of returns is maintained in Sections 4 and 5. Section 6 discusses modifying this null hypothesis to focus only on the mean of the return distribution. Section 7 presents analysis of the power of an event study. Section 8 presents nonparametric approaches to event studies which eliminate the need for parametric structure. In some cases theory provides hypotheses concerning the relation between the magnitude of the event abnormal return and firm characteristics. Section 9 presents a cross-sectional regression approach that is useful to investigate such hypotheses. Section 10 considers some further issues relating event study design and the paper closes with the concluding discussion in Section 11.

## *2. Procedure for an Event Study*

At the outset it is useful to briefly discuss the structure of an event study. This will provide a basis for the discussion of details later. While there is no unique structure, there is a general flow of analysis. This flow is discussed in this section.

The initial task of conducting an event study is to define the event of interest and identify the period over which the security prices of the firms involved in this event will be examined—the event window. For example, if one is looking at the information content of an earnings with daily data, the event will be the earnings announcement and the event window will include the one day of the announcement. It is customary to define the event window to be larger than the specific period of interest. This permits examination of periods surrounding the

event. In practice, the period of interest is often expanded to multiple days, including at least the day of the announcement and the day after the announcement. This captures the price effects of announcements which occur after the stock market closes on the announcement day. The periods prior to and after the event may also be of interest. For example, in the earnings announcement case, the market may acquire information about the earnings prior to the actual announcement and one can investigate this possibility by examining pre-event returns.

After identifying the event, it is necessary to determine the selection criteria for the inclusion of a given firm in the study. The criteria may involve restrictions imposed by data availability such as listing on the New York Stock Exchange or the American Stock Exchange or may involve restrictions such as membership in a specific industry. At this stage it is useful to summarize some sample characteristics (e.g., firm market capitalization, industry representation, distribution of events through time) and note any potential biases which may have been introduced through the sample selection.

Appraisal of the event's impact requires a measure of the abnormal return. The abnormal return is the actual ex post return of the security over the event window minus the normal return of the firm over the event window. The normal return is defined as the expected return without conditioning on the event taking place. For firm  $i$  and event date  $\tau$  the abnormal return is

$$AR_{i\tau} = R_{i\tau} - E(R_{i\tau}|X_\tau) \quad (1)$$

where  $AR_{i\tau}$ ,  $R_{i\tau}$ , and  $E(R_{i\tau}|X_\tau)$  are the abnormal, actual, and normal returns respectively for time period  $\tau$ .  $X_\tau$  is the information for the normal return. There are two common

choices for modeling the normal return—the *constant mean return model* where  $X_\tau$  is a constant, and the *market model* where  $X_\tau$  is the market return. The constant mean return model, as the name implies, assumes that the mean return of a given security is constant through time. The market model assumes a stable linear relation between the market return and the security return.

Given the selection of a normal performance model, the estimation window needs to be defined. The most common choice, when feasible, is using the period prior to the event window for the estimation window. For example, in an event study using daily data and the market model, the market model parameters could be estimated over the 120 days prior to the event. Generally the event period itself is not included in the estimation period to prevent the event from influencing the normal performance model parameter estimates.

With the parameter estimates for the normal performance model, the abnormal returns can be calculated. Next comes the design of the testing framework for the abnormal returns. Important considerations are defining the null hypothesis and determining the techniques for aggregating the individual firm abnormal returns.

The presentation of the empirical results follows the formulation of the econometric design. In addition to presenting the basic empirical results, the presentation of diagnostics can be fruitful. Occasionally, especially in studies with a limited number of event observations, the empirical results can be heavily influenced by one or two firms. Knowledge of this is important for gauging the importance of the results.

Ideally the empirical results will lead to insights relating to understanding the sources and causes of the effects (or lack

of effects) of the event under study. Additional analysis may be included to distinguish between competing explanations. Concluding comments complete the study.

### 3. *An Example of an Event Study*

The Financial Accounting Standards Board (FASB) and the Securities Exchange Commission strive to set reporting regulations so that financial statements and related information releases are informative about the value of the firm. In setting standards, the information content of the financial disclosures is of interest. Event studies provide an ideal tool for examining the information content of the disclosures.

In this section the description of an example selected to illustrate event study methodology is presented. One particular type of disclosure—quarterly earnings announcements—is considered. The objective is to investigate the information content of these announcements. In other words, the goal is to see if the release of accounting information provides information to the marketplace. If so there should be a correlation between the observed change of the market value of the company and the information.

The example will focus on the quarterly earnings announcements for the 30 firms in the Dow Jones Industrial Index over the five-year period from January 1989 to December 1993. These announcements correspond to the quarterly earnings for the last quarter of 1988 through the third quarter of 1993. The five years of data for 30 firms provide a total sample of 600 announcements. For each firm and quarter, three pieces of information are compiled: the date of the announcement, the actual earnings, and a measure of the expected earnings. The source of the date of the announcement

is Datastream, and the source of the actual earnings is Compustat.

If earnings announcements convey information to investors, one would expect the announcement impact on the market's valuation of the firm's equity to depend on the magnitude of the unexpected component of the announcement. Thus a measure of the deviation of the actual announced earnings from the market's prior expectation is required. For constructing such a measure, the mean quarterly earnings forecast reported by the Institutional Brokers Estimate System (I/B/E/S) is used to proxy for the market's expectation of earnings. I/B/E/S compiles forecasts from analysts for a large number of companies and reports summary statistics each month. The mean forecast is taken from the last month of the quarter. For example, the mean third quarter forecast from September 1990 is used as the measure of expected earnings for the third quarter of 1990.

To facilitate the examination of the impact of the earnings announcement on the value of the firm's equity, it is essential to posit the relation between the information release and the change in value of the equity. In this example the task is straightforward. If the earnings disclosures have information content, higher than expected earnings should be associated with increases in value of the equity and lower than expected earnings with decreases. To capture this association, each announcement is assigned to one of three categories: good news, no news, or bad news. Each announcement is categorized using the deviation of the actual earnings from the expected earnings. If the actual exceeds expected by more than 2.5 percent the announcement is designated as good news, and if the actual is more than 2.5 percent less than expected the announcement is designated as bad news. Those announce-

ment  
5 per  
pecte  
news.  
are g  
remai

With  
the n  
of the  
uity  
value  
speci  
an ev  
dow.  
to on  
used.  
ploys  
the e  
For e  
day p  
used  
prese  
study  
to ill

4

A  
to cal  
secur  
group  
and c  
gory  
conce  
and c  
gume  
ond  
cerni  
basec  
shoul  
econ  
sary  
the  
mode  
assu  
culat  
mal

ments where the actual earnings is in the 5 percent range centered about the expected earnings are designated as no news. Of the 600 announcements, 189 are good news, 173 are no news, and the remaining 238 are bad news.

With the announcements categorized, the next step is to specify the parameters of the empirical design to analyze the equity return, i.e., the percent change in value of the equity. It is necessary to specify a length of observation interval, an event window, and an estimation window. For this example the interval is set to one day, thus daily stock returns are used. A 41-day event window is employed, comprised of 20 pre-event days, the event day, and 20 post-event days. For each announcement the 250 trading day period prior to the event window is used as the estimation window. After presenting the methodology of an event study, this example will be drawn upon to illustrate the execution of a study.

#### 4. Models for Measuring Normal Performance

A number of approaches are available to calculate the normal return of a given security. The approaches can be loosely grouped into two categories—statistical and economic. Models in the first category follow from statistical assumptions concerning the behavior of asset returns and do not depend on any economic arguments. In contrast, models in the second category rely on assumptions concerning investors' behavior and are not based solely on statistical assumptions. It should, however, be noted that to use economic models in practice it is necessary to add statistical assumptions. Thus the potential advantage of economic models is not the absence of statistical assumptions, but the opportunity to calculate more precise measures of the normal return using economic restrictions.

For the statistical models, the assumption that asset returns are jointly multivariate normal and independently and identically distributed through time is imposed. This distributional assumption is sufficient for the constant mean return model and the market model to be correctly specified. While this assumption is strong, in practice it generally does not lead to problems because the assumption is empirically reasonable and inferences using the normal return models tend to be robust to deviations from the assumption. Also one can easily modify the statistical framework so that the analysis of the abnormal returns is autocorrelation and heteroskedasticity consistent by using a generalized method-of-moments approach.

##### A. Constant Mean Return Model

Let  $\mu_i$  be the mean return for asset  $i$ . Then the constant mean return model is

$$R_{it} = \mu_i + \zeta_{it} \quad (2)$$

$$E(\zeta_{it}) = 0 \quad \text{var}(\zeta_{it}) = \sigma_{\zeta_i}^2$$

where  $R_{it}$  is the period- $t$  return on security  $i$  and  $\zeta_{it}$  is the time period  $t$  disturbance term for security  $i$  with an expectation of zero and variance  $\sigma_{\zeta_i}^2$ .

Although the constant mean return model is perhaps the simplest model, Brown and Warner (1980, 1985) find it often yields results similar to those of more sophisticated models. This lack of sensitivity to the model can be attributed to the fact that the variance of the abnormal return is frequently not reduced much by choosing a more sophisticated model. When using daily data the model is typically applied to nominal returns. With monthly data the model can be applied to real returns or excess returns (the return in excess of the nominal risk free return generally measured using the U.S. Treasury Bill with one month to maturity) as well as nominal returns.



**B. Market Model**

The market model is a statistical model which relates the return of any given security to the return of the market portfolio. The model's linear specification follows from the assumed joint normality of asset returns. For any security  $i$  the market model is

$$R_{it} = \alpha_i + \beta_i R_{mt} + \epsilon_{it} \quad (3)$$

$$E(\epsilon_{it}) = 0 \quad \text{var}(\epsilon_{it}) = \sigma_{\epsilon_i}^2$$

where  $R_{it}$  and  $R_{mt}$  are the period- $t$  returns on security  $i$  and the market portfolio, respectively, and  $\epsilon_{it}$  is the zero mean disturbance term.  $\alpha_i$ ,  $\beta_i$ , and  $\sigma_{\epsilon_i}^2$  are the parameters of the market model. In applications a broad based stock index is used for the market portfolio, with the S&P 500 Index, the CRSP Value Weighted Index, and the CRSP Equal Weighted Index being popular choices.

The market model represents a potential improvement over the constant mean return model. By removing the portion of the return that is related to variation in the market's return, the variance of the abnormal return is reduced. This in turn can lead to increased ability to detect event effects. The benefit from using the market model will depend upon the  $R^2$  of the market model regression. The higher the  $R^2$  the greater is the variance reduction of the abnormal return, and the larger is the gain.

**C. Other Statistical Models**

A number of other statistical models have been proposed for modeling the normal return. A general type of statistical model is the *factor model*. Factor models are motivated by the benefits of reducing the variance of the abnormal return by explaining more of the variation in the normal return. Typically the factors are portfolios of traded securities.

The market model is an example of a one factor model. Other multifactor models include industry indexes in addition to the market. William Sharpe (1970) and Sharpe, Gordon Alexander, and Jeffery Bailey (1995, p. 303) provide discussion of index models with factors based on industry classification. Another variant of a factor model is a procedure which calculates the abnormal return by taking the difference between the actual return and a portfolio of firms of similar size, where size is measured by market value of equity. In this approach typically ten size groups are considered and the loading on the size portfolios is restricted to unity. This procedure implicitly assumes that expected return is directly related to market value of equity.

Generally, the gains from employing multifactor models for event studies are limited. The reason for the limited gains is the empirical fact that the marginal explanatory power of additional factors the market factor is small, and hence, there is little reduction in the variance of the abnormal return. The variance reduction will typically be greatest in cases where the sample firms have a common characteristic, for example they are all members of one industry or they are all firms concentrated in one market capitalization group. In these cases the use of a multifactor model warrants consideration.

The use of other models is dictated by data availability. An example of a normal performance return model implemented in situations with limited data is the market-adjusted return model. For some events it is not feasible to have a pre-event estimation period for the normal model parameters, and a market-adjusted abnormal return is used. The market-adjusted return model can be viewed as a restricted market model with  $\alpha_i$  constrained to be zero and  $\beta_i$  constrained to be one. Because the model coefficients

are prespecified, an estimation period is not required to obtain parameter estimates. An example of when such a model is used is in studies of the under pricing of initial public offerings. Jay Ritter (1991) presents such an example. A general recommendation is to only use such restricted models if necessary, and if necessary, consider the possibility of biases arising from the imposition of the restrictions.

#### D. Economic Models

Economic models can be cast as restrictions on the statistical models to provide more constrained normal return models. Two common economic models which provide restrictions are the Capital Asset Pricing Model (CAPM) and the Arbitrage Pricing Theory (APT). The CAPM due to Sharpe (1964) and John Lintner (1965) is an equilibrium theory where the expected return of a given asset is determined by its covariance with the market portfolio. The APT due to Stephen Ross (1976) is an asset pricing theory where the expected return of a given asset is a linear combination of multiple risk factors.

The use of the Capital Asset Pricing Model is common in event studies of the 1970s. However, deviations from the CAPM have been discovered, implying that the validity of the restrictions imposed by the CAPM on the market model is questionable.<sup>2</sup> This has introduced the possibility that the results of the studies may be sensitive to the specific CAPM restrictions. Because this potential for sensitivity can be avoided at little cost by using the market model, the use of the CAPM has almost ceased.

Similarly, other studies have employed multifactor normal performance models

motivated by the Arbitrage Pricing Theory. A general finding is that with the APT the most important factor behaves like a market factor and additional factors add relatively little explanatory power. Thus the gains from using an APT motivated model versus the market model are small. See Stephen Brown and Mark Weinstein (1985) for further discussion. The main potential gain from using a model based on the arbitrage pricing theory is to eliminate the biases introduced by using the CAPM. However, because the statistically motivated models also eliminate these biases, for event studies such models dominate.

#### 5. Measuring and Analyzing Abnormal Returns

In this section the problem of measuring and analyzing abnormal returns is considered. The framework is developed using the market model as the normal performance return model. The analysis is virtually identical for the constant mean return model.

Some notation is first defined to facilitate the measurement and analysis of abnormal returns. Returns will be indexed in event time using  $\tau$ . Defining  $\tau=0$  as the event date,  $\tau=T_1+1$  to  $\tau=T_2$  represents the event window, and  $\tau=T_0+1$  to  $\tau=T_1$  constitutes the estimation window. Let  $L_1=T_1-T_0$  and  $L_2=T_2-T_1$  be the length of the estimation window and the event window respectively. Even if the event being considered is an announcement on given date it is typical to set the event window length to be larger than one. This facilitates the use of abnormal returns around the event day in the analysis. When applicable, the post-event window will be from  $\tau=T_2+1$  to  $\tau=T_3$  and of length  $L_3=T_3-T_2$ . The timing sequence is illustrated with a time line in Figure 1.

<sup>2</sup>Eugene Fama and Kenneth French (1996) provide discussion of these anomalies.

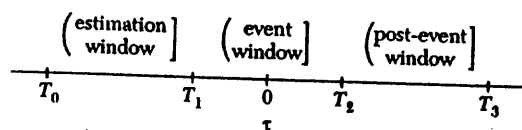


Figure 1. Time line for an event study.

It is typical for the estimation window and the event window not to overlap. This design provides estimators for the parameters of the normal return model which are not influenced by the returns around the event. Including the event window in the estimation of the normal model parameters could lead to the event returns having a large influence on the normal return measure. In this situation both the normal returns and the abnormal returns would capture the event impact. This would be problematic because the methodology is built around the assumption that the event impact is captured by the abnormal returns. On occasion, the post event window data is included with the estimation window data to estimate the normal return model. The goal of this approach is to increase the robustness of the normal market return measure to gradual changes in its parameters. In Section 6 expanding the null hypothesis to accommodate changes in the risk of a firm around the event is considered. In this case an estimation framework which uses the event window returns will be required.

#### A. Estimation of the Market Model

Under general conditions ordinary least squares (OLS) is a consistent estimation procedure for the market model parameters. Further, given the assumptions of Section 4, OLS is efficient. For the  $i^{\text{th}}$  firm in event time, the OLS estimators of the market model parameters for an estimation window of observations are

$$\hat{\beta}_i = \frac{\sum_{\tau=T_0+1}^{T_1} (R_{i\tau} - \hat{\mu}_i)(R_{m\tau} - \hat{\mu}_m)}{\sum_{\tau=T_0+1}^{T_1} (R_{m\tau} - \hat{\mu}_m)^2} \quad (4)$$

$$\hat{\alpha}_i = \hat{\mu}_i - \hat{\beta}_i \hat{\mu}_m \quad (5)$$

$$\hat{\sigma}_i^2 = \frac{1}{L_1 - 2} \sum_{\tau=T_0+1}^{T_1} (R_{i\tau} - \hat{\alpha}_i - \hat{\beta}_i R_{m\tau})^2 \quad (6)$$

where

$$\hat{\mu}_i = \frac{1}{L_1} \sum_{\tau=T_0+1}^{T_1} R_{i\tau}$$

$$\text{and } \hat{\mu}_m = \frac{1}{L_1} \sum_{\tau=T_0+1}^{T_1} R_{m\tau}$$

$R_{i\tau}$  and  $R_{m\tau}$  are the return in event period  $\tau$  for security  $i$  and the market respectively. The use of the OLS estimators to measure abnormal returns and to develop their statistical properties is addressed next. First, the properties of a given security are presented followed by consideration of the properties of abnormal returns aggregated across securities.

#### B. Statistical Properties of Abnormal Returns

Given the market model parameter estimates, one can measure and analyze the abnormal returns. Let  $AR_{i\tau}$ ,  $\tau = T_1 + 1, \dots, T_2$ , be the sample of  $L_2$  abnormal returns for firm  $i$  in the event window. Using the market model to measure the normal return, the sample abnormal return is

$$AR_{i\tau} = R_{i\tau} - \hat{\alpha}_i - \hat{\beta}_i R_{m\tau} \quad (7)$$

The abnormal return is the disturbance term of the market model calculated on an out of sample basis. Under the null hypothesis, conditional on the event win-



dow market returns, the abnormal returns will be jointly normally distributed with a zero conditional mean and conditional variance  $\sigma^2(AR_{it})$  where

$$\sigma^2(AR_{it}) = \sigma_{\varepsilon_i}^2 + \frac{1}{L_1} \left[ 1 + \frac{(R_{mt} - \hat{\mu}_m)^2}{\hat{\sigma}_m^2} \right] \quad (8)$$

From (8), the conditional variance has two components. One component is the disturbance variance  $\sigma_{\varepsilon_i}^2$  from (3) and a second component is additional variance due to the sampling error in  $\alpha_i$  and  $\beta_i$ . This sampling error, which is common for all the event window observations, also leads to serial correlation of the abnormal returns despite the fact that the true disturbances are independent through time. As the length of the estimation window  $L_1$  becomes large, the second term approaches zero as the sampling error of the parameters vanishes. The variance of the abnormal return will be  $\sigma_{\varepsilon_i}^2$  and the abnormal return observations will become independent through time. In practice, the estimation window can usually be chosen to be large enough to make it reasonable to assume that the contribution of the second component to the variance of the abnormal return is zero.

Under the null hypothesis,  $H_0$ , that the event has no impact on the behavior of returns (mean or variance) the distributional properties of the abnormal returns can be used to draw inferences over any period within the event window. Under  $H_0$  the distribution of the sample abnormal return of a given observation in the event window is

$$AR_{it} \sim N(0, \sigma^2(AR_{it})). \quad (9)$$

Next (9) is built upon to consider the aggregation of the abnormal returns.

### C. Aggregation of Abnormal Returns

The abnormal return observations must be aggregated in order to draw

overall inferences for the event of interest. The aggregation is along two dimensions—through time and across securities. We will first consider aggregation through time for an individual security and then will consider aggregation both across securities and through time. The concept of a cumulative abnormal return is necessary to accommodate a multiple period event window. Define  $CAR_i(\tau_1, \tau_2)$  as the sample cumulative abnormal return (CAR) from  $\tau_1$  to  $\tau_2$  where  $T_1 < \tau_1 \leq \tau_2 \leq T_2$ . The CAR from  $\tau_1$  to  $\tau_2$  is the sum of the included abnormal returns,

$$CAR_i(\tau_1, \tau_2) = \sum_{t=\tau_1}^{\tau_2} AR_{it} \quad (10)$$

Asymptotically (as  $L_1$  increases) the variance of  $CAR_i$  is

$$\sigma_i^2(\tau_1, \tau_2) = (\tau_2 - \tau_1 + 1) \sigma_{\varepsilon_i}^2 \quad (11)$$

This large sample estimator of the variance can be used for reasonable values of  $L_1$ . However, for small values of  $L_1$  the variance of the cumulative abnormal return should be adjusted for the effects of the estimation error in the normal model parameters. This adjustment involves the second term of (8) and a further related adjustment for the serial covariance of the abnormal return.

The distribution of the cumulative abnormal return under  $H_0$  is

$$CAR_i(\tau_1, \tau_2) \sim N(0, \sigma_i^2(\tau_1, \tau_2)). \quad (12)$$

Given the null distributions of the abnormal return and the cumulative abnormal return, tests of the null hypothesis can be conducted.

However, tests with one event observation are not likely to be useful so it is necessary to aggregate. The abnormal return observations must be aggregated for the event window and across observations of the event. For this aggregation,

TABLE 1

Event Day	Market Model					
	Good News		No News		Bad News	
	AR	CAR	AR	CAR	AR	CAR
-20	.093	.093	.080	.080	-.107	-.107
-19	-.177	-.084	.018	.098	-.180	-.286
-18	.088	.004	.012	.110	.029	-.258
-17	.024	.029	-.151	-.041	-.079	-.337
-16	-.018	.011	-.019	-.060	-.010	-.346
-15	-.040	-.029	.013	-.047	-.054	-.401
-14	.038	.008	.040	-.007	-.021	-.421
-13	.056	.064	-.057	-.065	.007	-.414
-12	.065	.129	.146	.081	-.090	-.504
-11	.069	.199	-.020	.061	-.088	-.592
-10	.028	.227	.025	.087	-.092	-.683
-9	.155	.382	.115	.202	-.040	-.724
-8	.057	.438	.070	.272	.072	-.652
-7	-.010	.428	-.106	.166	-.026	-.677
-6	.104	.532	.026	.192	-.013	-.690
-5	.085	.616	-.085	.107	.164	-.527
-4	.099	.715	.040	.147	-.139	-.666
-3	.117	.832	.036	.183	.098	-.568
-2	.006	.838	.226	.409	-.112	-.680
-1	.164	1.001	-.168	.241	-.180	-.860
0	.965	1.966	-.091	.150	-.679	-1.539
1	.251	2.217	-.008	.142	-.204	-1.743
2	-.014	2.203	.007	.148	.072	-1.672
3	-.164	2.039	.042	.190	.083	-1.589
4	-.014	2.024	.000	.190	.106	-1.483
5	.135	2.160	-.038	.152	.194	-1.289
6	-.052	2.107	-.302	-.150	.076	-1.213
7	.060	2.167	-.199	-.349	.120	-1.093
8	.155	2.323	-.108	-.457	-.041	-1.134
9	-.008	2.315	-.146	-.603	-.069	-1.203
10	.164	2.479	.082	-.521	.130	-1.073
11	-.081	2.398	.040	-.481	-.009	-1.082
12	-.058	2.341	.246	-.235	-.038	-1.119
13	-.165	2.176	.014	-.222	.071	-1.048
14	-.081	2.095	-.091	-.312	.019	-1.029
15	-.007	2.088	-.001	-.314	-.043	-1.072
16	.065	2.153	-.020	-.334	-.086	-1.159
17	.081	2.234	.017	-.317	-.050	-1.208
18	.172	2.406	.054	-.263	.066	-1.142
19	-.043	2.363	.119	-.144	-.088	-1.230
20	.013	2.377	.094	-.050	-.028	-1.258

## MacKinlay: Event Studies in Economics and Finance

23

TABLE 1 (Cont.)

Constant Mean Return Model					
Good News		No News		Bad News	
AR	CAR	AR	CAR	AR	CAR
.105	.105	.019	.019	-.077	-.077
-.235	-.129	-.048	-.029	-.142	-.219
.069	-.060	-.086	-.115	-.043	-.262
-.026	-.086	-.140	-.255	-.057	-.319
-.086	-.172	.039	-.216	-.075	-.394
-.183	-.355	.099	-.117	-.037	-.431
-.020	-.375	-.150	-.266	-.101	-.532
-.025	-.399	-.191	-.458	-.069	-.601
.101	-.298	.133	-.325	-.106	-.707
.126	-.172	.006	-.319	-.169	-.876
.134	-.038	.103	-.216	-.009	-.885
.210	.172	.022	-.194	.011	-.874
.106	.278	.163	-.031	.135	-.738
-.002	.277	.009	-.022	-.027	-.765
.011	.288	-.029	-.051	.030	-.735
.061	.349	-.068	-.120	.320	-.415
.031	.379	.089	-.031	-.205	-.620
.067	.447	.013	-.018	.085	-.536
.010	.456	.311	.294	-.256	-.791
.198	.654	-.170	.124	-.227	-1.018
1.034	1.688	-.164	-.040	-.643	-1.661
.357	2.045	-.170	-.210	-.212	-1.873
-.013	2.033	.054	-.156	.078	-1.795
.088	1.944	-.121	-.277	.146	-1.648
.041	1.985	.023	-.253	.149	-1.499
.248	2.233	-.003	-.256	.286	-1.214
-.035	2.198	-.319	-.575	.070	-1.143
.017	2.215	-.112	-.687	.102	-1.041
.112	2.326	-.187	-.874	.056	-.986
-.052	2.274	-.057	-.931	-.071	-1.056
.147	2.421	.203	-.728	.267	-.789
-.013	2.407	.045	-.683	.006	-.783
-.054	2.354	.299	-.384	.017	-.766
-.246	2.107	-.067	-.451	.114	-.652
-.011	2.096	-.024	-.475	.089	-.564
-.027	2.068	-.059	-.534	-.022	-.585
.103	2.171	-.046	-.580	-.084	-.670
.066	2.237	-.098	-.677	-.054	-.724
.110	2.347	.021	-.656	-.071	-.795
-.055	2.292	.088	-.568	.026	-.769
.019	2.311	.013	-.554	-.115	-.884

Abnormal returns for an event study of the information content of earnings announcements. The sample consists of a total of 600 quarterly announcements for the 30 companies in the Dow Jones Industrial Index for the five year period January 1989 to December 1993. Two models are considered for the normal returns, the market model using the CRSP value-weighted index and the constant return model. The announcements are categorized into three groups, good news, no news, and bad news. AR is the sample average abnormal return for the specified day in event time and CAR is the sample average cumulative abnormal return for day -20 to the specified day. Event time is days relative to the announcement date.

it is assumed that there is not any clustering. That is, there is not any overlap in the event windows of the included securities. The absence of any overlap and the maintained distributional assumptions imply that the abnormal returns and the cumulative abnormal returns will be independent across securities. Later inferences with clustering will be discussed.

The individual securities' abnormal returns can be aggregated using  $AR_{it}$  from (7) for each event period,  $\tau = T_1 + 1, \dots, T_2$ . Given  $N$  events, the sample aggregated abnormal returns for period  $\tau$  is

$$\bar{AR}_\tau = \frac{1}{N} \sum_{i=1}^N AR_{it} \quad (13)$$

and for large  $L_1$ , its variance is

$$\text{var}(\bar{AR}_\tau) = \frac{1}{N^2} \sum_{i=1}^N \sigma_{\epsilon_i}^2 \quad (14)$$

Using these estimates, the abnormal returns for any event period can be analyzed.

The average abnormal returns can then be aggregated over the event window using the same approach as that used to calculate the cumulative abnormal return for each security  $i$ . For any interval in the event window

$$\overline{CAR}(\tau_1, \tau_2) = \sum_{\tau=\tau_1}^{\tau_2} \bar{AR}_\tau \quad (15)$$

$$\text{var}(\overline{CAR}(\tau_1, \tau_2)) = \sum_{\tau=\tau_1}^{\tau_2} \text{var}(\bar{AR}_\tau) \quad (16)$$

Observe that equivalently one can form the CAR's security by security and then aggregate through time,

$$\overline{CAR}(\tau_1, \tau_2) = \frac{1}{N} \sum_{i=1}^N CAR_i(\tau_1, \tau_2) \quad (17)$$

$$\text{var}(\overline{CAR}(\tau_1, \tau_2)) = \frac{1}{N^2} \sum_{i=1}^N \sigma_i^2(\tau_1, \tau_2) \quad (18)$$

For the variance estimators the assumption that the event windows of the  $N$  securities do not overlap is used to set the covariance terms to zero. Inferences about the cumulative abnormal returns can be drawn using

$$\overline{CAR}(\tau_1, \tau_2) \sim N[0, \text{var}(\overline{CAR}(\tau_1, \tau_2))] \quad (19)$$

to test the null hypothesis that the abnormal returns are zero. In practice, because  $\sigma_{\epsilon_i}^2$  is unknown, an estimator must be used to calculate the variance of the abnormal returns as in (14). The usual sample variance measure of  $\sigma_{\epsilon_i}^2$  from the market model regression in the estimation window is an appropriate choice. Using this to calculate  $\text{var}(\bar{AR}_\tau)$  in (14),  $H_0$  can be tested using

$$\theta_1 = \frac{\overline{CAR}(\tau_1, \tau_2)}{\text{var}(\overline{CAR}(\tau_1, \tau_2))^{1/2}} \sim N(0, 1) \quad (20)$$

This distributional result is asymptotic with respect to the number of securities  $N$  and the length of estimation window  $L_1$ .

Modifications to the basic approach presented above are possible. One common modification is to standardize each abnormal return using an estimator of its standard deviation. For certain alternatives, such standardization can lead to more powerful tests. James Patell (1976) presents tests based on standardization and Brown and Warner (1980, 1985) provide comparisons with the basic approach.

#### D. CAR's for the Earnings Announcement Example

The information content of earnings example previously described illustrates the use of sample abnormal residuals and sample cumulative abnormal returns. Table 1 presents the abnormal returns av-

0.025

0.02

0.015

0.01

0.005

CAR 0

-0.005

-0.01

-0.015

-0.02

-0.025

Fig  
day

erage  
(30  
as w  
norm  
ings  
mod  
mod  
mea  
tive  
with  
in 1  
con.  
2b.

T  
con  
the  
evi  
sis

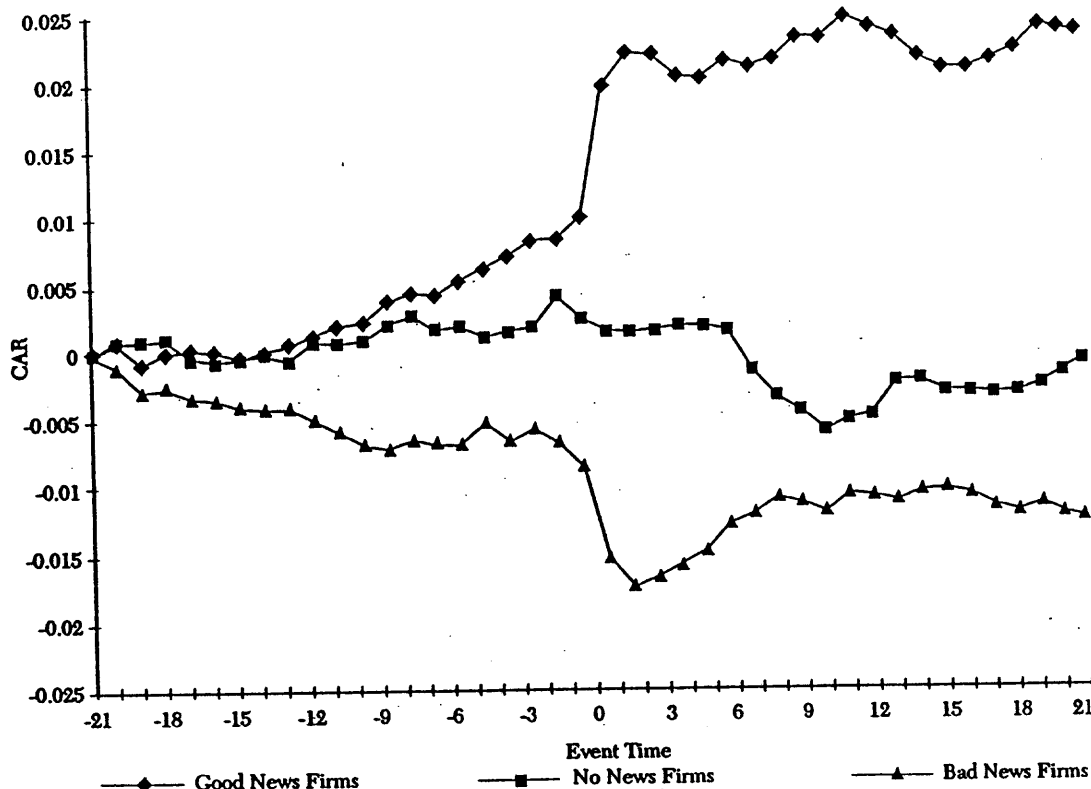


Figure 2a. Plot of cumulative abnormal return for earning announcements from event day -20 to event day 20. The abnormal return is calculated using the market model as the normal return measure.

eraged across the 600 event observations (30 firms, 20 announcements per firm) as well as the aggregated cumulative abnormal return for each of the three earnings news categories. Two normal return models are considered; the market model and for comparison, the constant mean return model. Plots of the cumulative abnormal returns are also included, with the CAR's from the market model in Figure 2a and the CAR's from the constant mean return model in Figure 2b.

The results of this example are largely consistent with the existing literature on the information content of earnings. The evidence strongly supports the hypothesis that earnings announcements do in-

deed convey information useful for the valuation of firms. Focusing on the announcement day (day 0) the sample average abnormal return for the good news firm using the market model is 0.965 percent. Given the standard error of the one day good news average abnormal return is 0.104 percent, the value of  $\theta_1$  is 9.28 and the null hypothesis that the event has no impact is strongly rejected. The story is the same for the bad news firms. The event day sample abnormal return is -0.679 percent, with a standard error of 0.098 percent, leading to  $\theta_1$  equal to -6.93 and again strong evidence against the null hypothesis. As would be expected, the abnormal return of the no news firms is small at -0.091 percent and

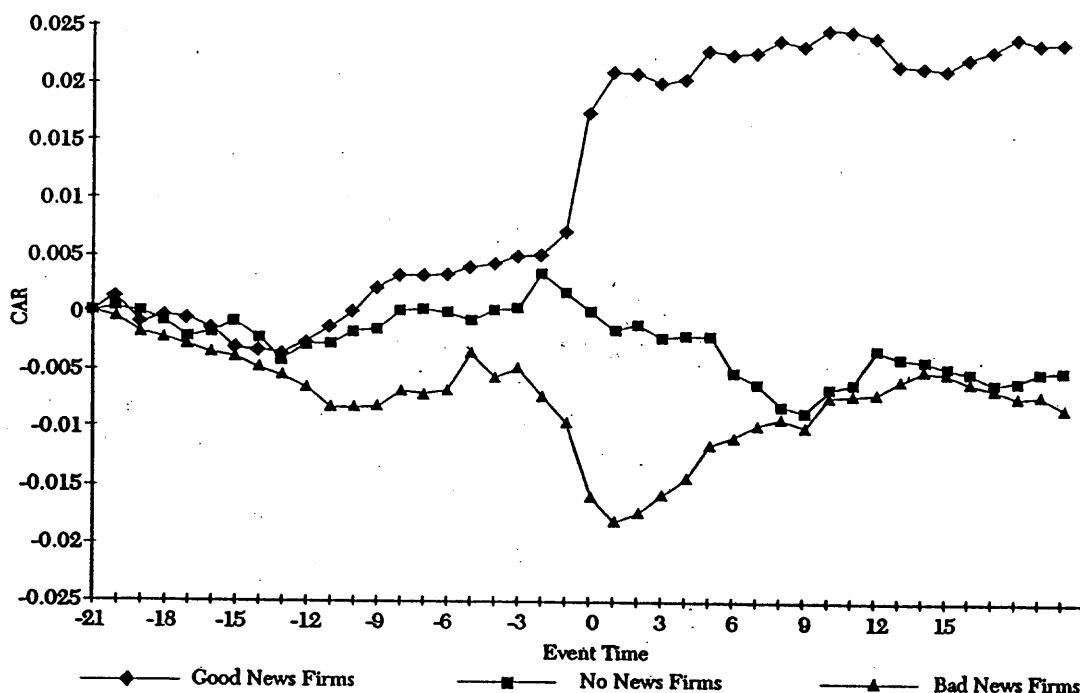


Figure 2b. Plot of cumulative abnormal return for earning announcements from event day -20 to event day 20. The abnormal return is calculated using the constant mean return model as the normal return

with a standard error of 0.098 percent is less than one standard error from zero. There is some evidence of the announcement effect on day one. The average abnormal return is 0.251 percent and -0.204 percent for the good news and the bad news firms respectively. Both these values are more than two standard errors from zero. The source of these day one effects is likely to be that some of the earnings announcements are made on event day zero after the close of the stock market. In these cases, the effects will be captured in the return on day one.

The conclusions using the abnormal returns from the constant return model are consistent with those from the market model. However, there is some loss of precision using the constant return model, as the variance of the average abnormal return increases for all three

categories. When measuring abnormal returns with the constant mean return model the standard errors increase from 0.104 percent to 0.130 percent for good news firms, from 0.098 percent to 0.124 percent for no news firms, and from 0.098 percent to 0.131 percent for bad news firms. These increases are to be expected when considering a sample of large firms such as those in the Dow Index because these stocks tend to have an important market component whose variability is eliminated using the market model.

The CAR plots show that to some extent the market gradually learns about the forthcoming announcement. The average CAR of the good news firms gradually drifts up in days -20 to -1 and the average CAR of the bad news firms gradually drifts down over this period. In the days after the an-

nou  
as  
doe  
tica  
bad  
eigl  
E.

T  
turn  
dov  
ove  
tion  
the  
ma  
cov  
the  
wir  
bet  
be  
sen  
tur  
Be  
pro  
(  
two  
ag  
eve  
of  
Th  
lat

is  
ou  
in  
in  
se  
is  
to  
or  
T  
a  
d  
aj  
K  
(  
V  
tl



nouncement the CAR is relatively stable as would be expected, although there does tend to be a slight (but statistically insignificant) increase with the bad news firms in days two through eight.

#### E. *Inferences with Clustering*

The analysis aggregating abnormal returns has assumed that the event windows of the included securities do not overlap in calendar time. This assumption allows us to calculate the variance of the aggregated sample cumulative abnormal returns without concern about the covariances across securities because they are zero. However, when the event windows do overlap and the covariances between the abnormal returns will not be zero, the distributional results presented for the aggregated abnormal returns are no longer applicable. Victor Bernard (1987) discusses some of the problems related to clustering.

Clustering can be accommodated in two ways. The abnormal returns can be aggregated into a portfolio dated using event time and the security level analysis of Section 5 can be applied to the portfolio. This approach will allow for cross correlation of the abnormal returns.

A second method to handle clustering is to analyze the abnormal returns without aggregation. One can consider testing the null hypothesis of the event having no impact using unaggregated security by security data. This approach is applied most commonly when there is total clustering, that is, there is an event on the same day for a number of firms. The basic approach is an application of a multivariate regression model with dummy variables for the event date. This approach is developed in the papers of Katherine Schipper and Rex Thompson (1983, 1985) and Daniel Collins and Warren Dent (1984). The advantage of the approach is that, unlike the portfolio

approach, an alternative hypothesis where some of the firms have positive abnormal returns and some of the firms have negative abnormal returns can be accommodated. However, in general the approach has two drawbacks—frequently the test statistic will have poor finite sample properties except in special cases and often the test will have little power against economically reasonable alternatives. The multivariate framework and its analysis is similar to the analysis of multivariate tests of asset pricing models. MacKinlay (1987) provides analysis in that context.

#### 6. *Modifying the Null Hypothesis*

Thus far the focus has been on a single null hypothesis—that the given event has no impact on the behavior of the returns. With this null hypothesis either a mean effect or a variance effect will represent a violation. However, in some applications one may be interested in testing for a mean effect. In these cases, it is necessary to expand the null hypothesis to allow for changing (usually increasing) variances. To allow for changing variance as part of the null hypothesis, it is necessary to eliminate the reliance on the past returns to estimate the variance of the aggregated cumulative abnormal returns. This is accomplished by using the cross section of cumulative abnormal returns to form an estimator of the variance for testing the null hypothesis. Ekkehart Boehmer, Jim Musumeci, and Annette Poulsen (1991) discuss methodology to accommodate changing variance.

The cross sectional approach to estimating the variance can be applied to the average cumulative abnormal return ( $\overline{CAR}(\tau_1, \tau_2)$ ). Using the cross-section to form an estimator of the variance gives



$$\begin{aligned} \text{var}(\overline{CAR}(\tau_1, \tau_2)) \\ = \frac{1}{N^2} \sum_{i=1}^N (CAR_i(\tau_1, \tau_2) \\ - \overline{CAR}(\tau_1, \tau_2))^2. \quad (21) \end{aligned}$$

For this estimator of the variance to be consistent, the abnormal returns need to be uncorrelated in the cross-section. An absence of clustering is sufficient for this requirement. Note that cross-sectional homoskedasticity is not required. Given this variance estimator, the null hypothesis that the cumulative abnormal returns are zero can then be tested using the usual theory.

One may also be interested in the question of the impact of an event on the risk of a firm. The relevant measure of risk must be defined before this question can be addressed. One choice as a risk measure is the market model beta which is consistent with the Capital Asset Pricing Model being appropriate. Given this choice, the market model can be formulated to allow the beta to change over the event window and the stability of the risk can be examined. Edward Kane and Haluk Unal (1988) present an application of this idea.

### 7. Analysis of Power

An important consideration when setting up an event study is the ability to detect the presence of a non-zero abnormal return. The inability to distinguish between the null hypothesis and economically interesting alternatives would suggest the need for modification of the design. In this section the question of the likelihood of rejecting the null hypothesis for a specified level of abnormal return associated with an event is addressed. Formally, the power of the test is evaluated.

Consider a two-sided test of the null hypothesis using the cumulative abnormal return based statistic  $\theta_1$  from (20). It is assumed that the abnormal returns are uncorrelated across securities; thus

the variance of  $\overline{CAR}$  is  $1/N^2 \sum_{i=1}^N \sigma_i^2(\tau_1, \tau_2)$

and  $N$  is the sample size. Because the null distribution of  $\theta_1$  is standard normal, for a two sided test of size  $\alpha$ , the null hypothesis will be rejected if  $\theta_1$  is in the critical region, that is,

$$\theta_1 < c\left(\frac{\alpha}{2}\right) \text{ or } \theta_1 > c\left(1 - \frac{\alpha}{2}\right)$$

where  $c(x) = \Phi^{-1}(x)$ .  $\Phi(\cdot)$  is the standard normal cumulative distribution function (CDF).

Given the specification of the alternative hypothesis  $H_A$  and the distribution of  $\theta_1$  for this alternative, the power of a test of size  $\alpha$  can be tabulated using the power function,

$$\begin{aligned} P(\alpha, H_A) = & pr\left(\theta_1 < c\left(\frac{\alpha}{2}\right) | H_A\right) \\ & + pr\left(\theta_1 > c\left(1 - \frac{\alpha}{2}\right) | H_A\right). \quad (22) \end{aligned}$$

The distribution of  $\theta_1$  under the alternative hypothesis considered below will be normal. The mean will be equal to the true cumulative abnormal return divided by the standard deviation of  $\overline{CAR}$  and the variance will be equal to one.

To tabulate the power one must posit economically plausible scenarios. The alternative hypotheses considered are four levels of abnormal returns, 0.5 percent, 1.0 percent, 1.5 percent, and 2.0 percent and two levels of the average variance for the cumulative abnormal return of a given security over the event period, 0.0004 and 0.0016. The

TABLE 2

Sample Size	Abnormal Return				Abnormal Return			
	.005	.010	.015	.020	.005	.010	.015	.020
	$\sigma = 0.02$				$\sigma = 0.04$			
1	0.06	0.08	0.12	0.17	0.05	0.06	0.07	0.08
2	0.06	0.11	0.19	0.29	0.05	0.06	0.08	0.11
3	0.07	0.14	0.25	0.41	0.06	0.07	0.10	0.14
4	0.08	0.17	0.32	0.52	0.06	0.08	0.12	0.17
5	0.09	0.20	0.39	0.61	0.06	0.09	0.13	0.20
6	0.09	0.23	0.45	0.69	0.06	0.09	0.15	0.23
7	0.10	0.26	0.51	0.75	0.06	0.10	0.17	0.26
8	0.11	0.29	0.56	0.81	0.06	0.11	0.19	0.29
9	0.12	0.32	0.61	0.85	0.07	0.12	0.20	0.32
10	0.12	0.35	0.66	0.89	0.07	0.12	0.22	0.35
11	0.13	0.38	0.70	0.91	0.07	0.13	0.24	0.38
12	0.14	0.41	0.74	0.93	0.07	0.14	0.25	0.41
13	0.15	0.44	0.77	0.95	0.07	0.15	0.27	0.44
14	0.15	0.46	0.80	0.96	0.08	0.15	0.29	0.46
15	0.16	0.49	0.83	0.97	0.08	0.16	0.31	0.49
16	0.17	0.52	0.85	0.98	0.08	0.17	0.32	0.52
17	0.18	0.54	0.87	0.98	0.08	0.18	0.34	0.54
18	0.19	0.56	0.89	0.99	0.08	0.19	0.36	0.56
19	0.19	0.59	0.90	0.99	0.08	0.19	0.37	0.59
20	0.20	0.61	0.92	0.99	0.09	0.20	0.39	0.61
25	0.24	0.71	0.96	1.00	0.10	0.24	0.47	0.71
30	0.28	0.78	0.98	1.00	0.11	0.28	0.54	0.78
35	0.32	0.84	0.99	1.00	0.11	0.32	0.60	0.84
40	0.35	0.89	1.00	1.00	0.12	0.35	0.66	0.89
45	0.39	0.92	1.00	1.00	0.13	0.39	0.71	0.92
50	0.42	0.94	1.00	1.00	0.14	0.42	0.76	0.94
60	0.49	0.97	1.00	1.00	0.16	0.49	0.83	0.97
70	0.55	0.99	1.00	1.00	0.18	0.55	0.88	0.99
80	0.61	0.99	1.00	1.00	0.20	0.61	0.92	0.99
90	0.66	1.00	1.00	1.00	0.22	0.66	0.94	1.00
100	0.71	1.00	1.00	1.00	0.24	0.71	0.96	1.00
120	0.78	1.00	1.00	1.00	0.28	0.78	0.98	1.00
140	0.84	1.00	1.00	1.00	0.32	0.84	0.99	1.00
160	0.89	1.00	1.00	1.00	0.35	0.89	1.00	1.00
180	0.92	1.00	1.00	1.00	0.39	0.92	1.00	1.00
200	0.94	1.00	1.00	1.00	0.42	0.94	1.00	1.00

Power of event study methodology for test of the null hypothesis that the abnormal return is zero. The power is reported for a two-sided test using  $\theta_1$  with a size of 5 percent. The sample size is the number of event observations included the study and  $\sigma$  is the square root of the average variance of the abnormal return across firms.

sample size, that is the number of securities for which the event occurs, is varied from one to 200. The power for a test with a size of 5 percent is documented. With  $\alpha = 0.05$ , the critical val-

ues calculated using  $c(\alpha/2)$  and  $c(1 - \alpha/2)$  are -1.96 and 1.96 respectively. Of course, in applications, the power of the test should be considered when selecting the size.

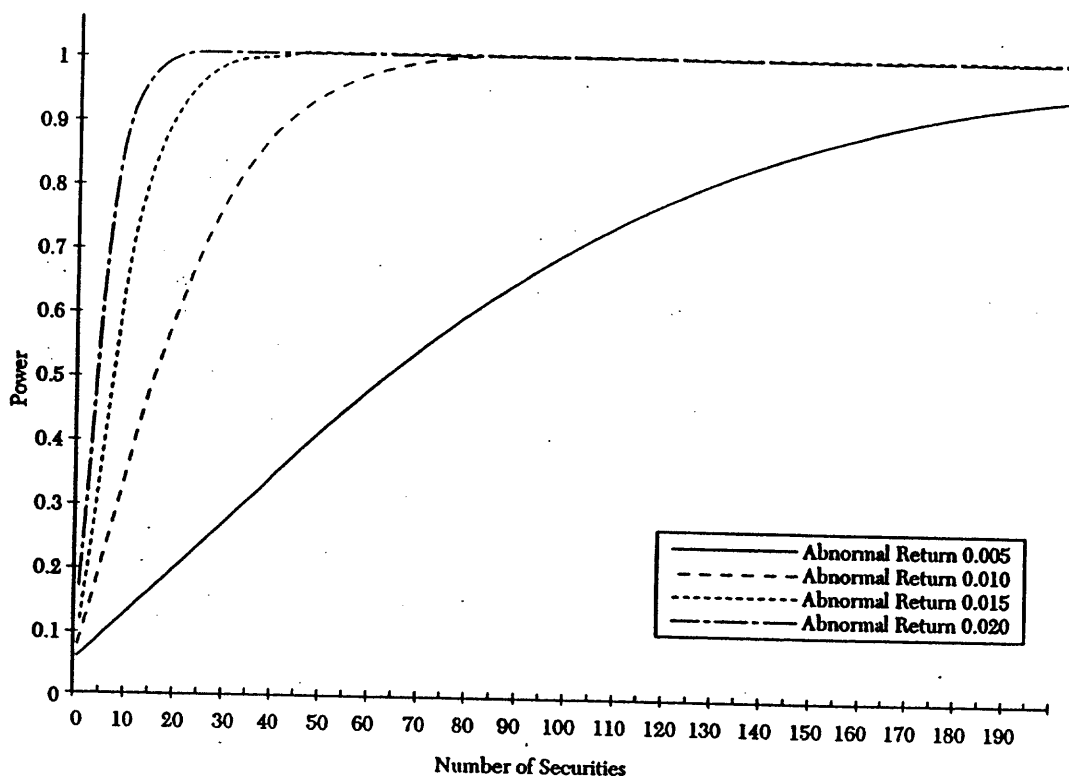


Figure 3a. Power of event study test statistic  $\theta_1$  to reject the null hypothesis that the abnormal return is zero, when the square root of the average variance of the abnormal return across firms is 2 percent.

The power results are presented in Table 2, and are plotted in Figures 3a and 3b. The results in the left panel of Table 2 and Figure 3a are for the case where the average variance is 0.0004. This corresponds to a cumulative abnormal return standard deviation of 2 percent and is an appropriate value for an event which does not lead to increased variance and can be examined using a one-day event window. In terms of having high power this is the best case scenario. The results illustrate that when the abnormal return is only 0.5 percent the power can be low. For example with a sample size of 20 the power of a 5 percent test is only 0.20. One needs a sample of over 60 firms before the power reaches 0.50. However, for a given sample size, increases in power

are substantial when the abnormal return is larger. For example, when the abnormal return is 2.0 percent the power of a 5 percent test with 20 firms is almost 1.00 with a value of 0.99. The general results for a variance of 0.0004 is that when the abnormal return is larger than 1 percent the power is quite high even for small sample sizes. When the abnormal return is small a larger sample size is necessary to achieve high power.

In the right panel of Table 2 and in Figure 3b the power results are presented for the case where the average variance of the cumulative abnormal return is 0.0016. This case corresponds roughly to either a multi-day event window or to a one-day event window with the event leading to increased variance

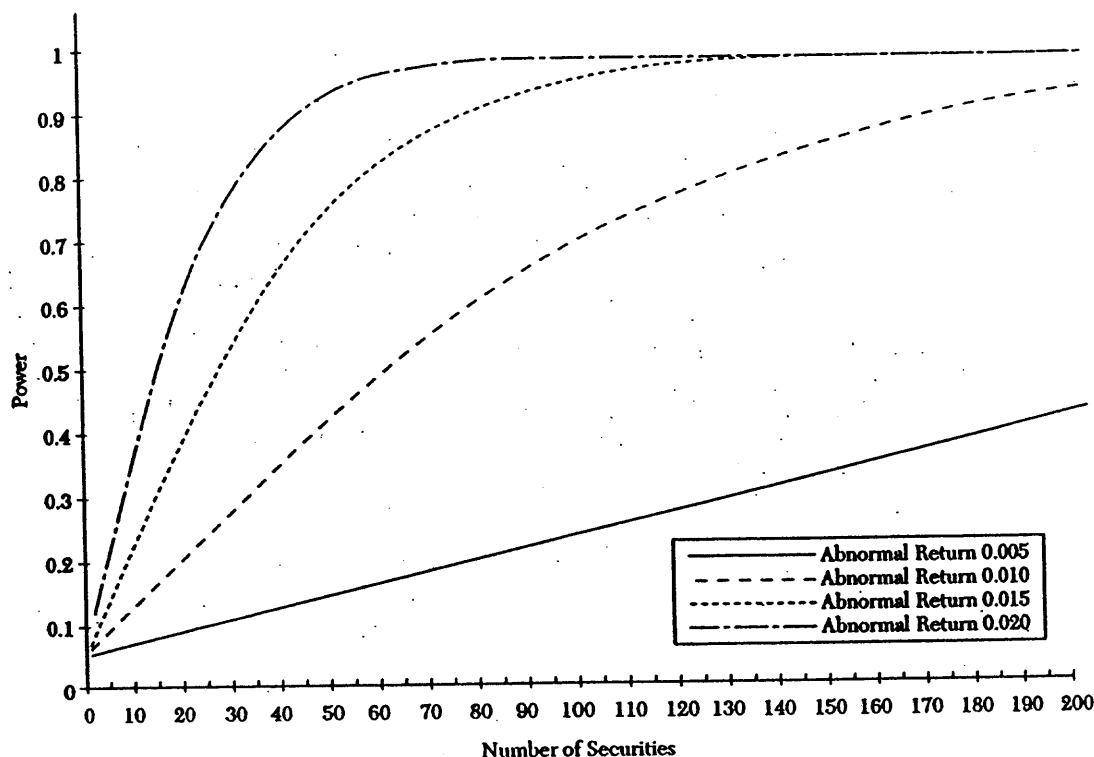


Figure 3b. Power of event study test statistic  $\theta_1$  to reject the null hypothesis that the abnormal return is zero, when the square root of the average variance of the abnormal return across firms is 4 percent.

which is accommodated as part of the null hypothesis. When the average variance of the CAR is increased from 0.0004 to 0.0016 there is a dramatic power decline for a 5 percent test. When the CAR is 0.5 percent the power is only 0.09 with 20 firms and is only 0.42 with a sample of 200 firms. This magnitude of abnormal return is difficult to detect with the larger variance. In contrast, when the CAR is as large as 1.5 percent or 2.0 percent the 5 percent test is still has reasonable power. For example, when the abnormal return is 1.5 percent and there is a sample size of 30 the power is 0.54. Generally if the abnormal return is large one will have little difficulty rejecting the null hypothesis of no abnormal return.

In the preceding analysis the power is

considered analytically for the given distributional assumptions. If the distributional assumptions are inappropriate then the results may differ. However, Brown and Warner (1985) consider this possible difference and find that the analytical computations and the empirical power are very close.

It is difficult to make general conclusions concerning the adequacy of the ability of event study methodology to detect non-zero abnormal returns. When conducting an event study it is best to evaluate the power given the parameters and objectives of the study. If the power seems sufficient then one can proceed, otherwise one should search for ways of increasing the power. This can be done by increasing the sample size, shortening the event window, or by

developing more specific predictions to test.

### 8. Nonparametric Tests

The methods discussed to this point are parametric in nature, in that specific assumptions have been made about the distribution of abnormal returns. Alternative approaches are available which are nonparametric in nature. These approaches are free of specific assumptions concerning the distribution of returns. Common nonparametric tests for event studies are the sign test and the rank test. These tests are discussed next.

The sign test, which is based on the sign of the abnormal return, requires that the abnormal returns (or more generally cumulative abnormal returns) are independent across securities and that the expected proportion of positive abnormal returns under the null hypothesis is 0.5. The basis of the test is that, under the null hypothesis, it is equally probable that the CAR will be positive or negative. If, for example, the null hypothesis is that there is a positive abnormal return associated with a given event, the null hypothesis is  $H_0: p \leq 0.5$  and the alternative is  $H_A: p > 0.5$  where  $p = pr[CAR_t \geq 0.0]$ . To calculate the test statistic we need the number of cases where the abnormal return is positive,  $N^+$ , and the total number of cases,  $N$ . Letting  $\theta_2$  be the test statistic,

$$\theta_2 = \left[ \frac{N^+}{N} - 0.5 \right] \frac{\sqrt{N}}{0.5} \sim N(0,1). \quad (23)$$

This distributional result is asymptotic. For a test of size  $(1 - \alpha)$ ,  $H_0$  is rejected if  $\theta_2 > \Phi^{-1}(\alpha)$ .

A weakness of the sign test is that it may not be well specified if the distribution of abnormal returns is skewed as can be the case with daily data. In response to this possible shortcoming,

Charles Corrado (1989) proposes a nonparametric rank test for abnormal performance in event studies. A brief description of his test of no abnormal return for event day zero follows. The framework can be easily altered for more general tests.

Drawing on notation previously introduced, consider a sample of  $L_2$  abnormal returns for each of  $N$  securities. To implement the rank test, for each security it is necessary to rank the abnormal returns from one to  $L_2$ . Define  $K_{it}$  as the rank of the abnormal return of security  $i$  for event time period  $\tau$ . Recall,  $\tau$  ranges from  $T_1 + 1$  to  $T_2$  and  $\tau = 0$  is the event day. The rank test uses the fact that the expected rank of the event day is  $(L_2 + 1)/2$  under the null hypothesis. The test statistic for the null hypothesis of no abnormal return on event day zero is

$$\theta_3 = \frac{1}{N} \sum_{i=1}^N \left( K_{i0} - \frac{L_2 + 1}{2} \right) / s(K) \quad (24)$$

where

$$s(K) = \sqrt{\frac{1}{L_2} \sum_{\tau=T_1+1}^{T_2} \left( \frac{1}{N} \sum_{i=1}^N \left( K_{i\tau} - \frac{L_2 + 1}{2} \right)^2 \right)}. \quad (25)$$

Tests of the null hypothesis can be implemented using the result that the asymptotic null distribution of  $\theta_3$  is standard normal. Corrado (1989) includes further discussion of details of this test.

Typically, these nonparametric tests are not used in isolation but in conjunction with the parametric counterparts. Inclusion of the nonparametric tests provides a check of the robustness of conclusions based on parametric tests. Such a check can be worthwhile as illustrated by the work of Cynthia Campbell and Charles Wasley (1993). They find that for NASDAQ stocks daily returns the nonparametric rank test provides more reliable inferences than do the standard parametric tests.

The exami  
magnit  
charac  
servati  
helpful  
for the  
cross-s  
appropri  
cross-s  
mal re  
terest.

Give  
observ:  
regress

$$AR_j =$$

where  
servatic  
acterist  
the zer  
uncorre  
M are  
regress  
OLS.  
tionally  
inferen  
usual C  
without  
erosked  
ing sta  
the ap  
The us  
standar  
there is  
of (26)

Paul  
(1986)  
section  
tive ab  
ment c  
on the  
age of  
firm ar



### 9. Cross-Sectional Models

Theoretical insights can result from examining the association between the magnitude of the abnormal return and characteristics specific to the event observation. Often such an exercise can be helpful when multiple hypotheses exist for the source of the abnormal return. A cross-sectional regression model is an appropriate tool to investigate this association. The basic approach is to run a cross-sectional regression of the abnormal returns on the characteristics of interest.

Given a sample of  $N$  abnormal return observations and  $M$  characteristics, the regression model is:

$$AR_j = \delta_0 + \delta_1 x_{1j} + \dots + \delta_M x_{Mj} + \eta_j \quad (26)$$

$$E(\eta_j) = 0 \quad (27)$$

where  $AR_j$  is the  $j^{\text{th}}$  abnormal return observation,  $x_{mj}, m = 1, \dots, M$ , are  $M$  characteristics for the  $j^{\text{th}}$  observation and  $\eta_j$  is the zero mean disturbance term that is uncorrelated with the  $x$ 's.  $\delta_m, m = 0, \dots, M$  are the regression coefficients. The regression model can be estimated using OLS. Assuming the  $\eta_j$ 's are cross-sectionally uncorrelated and homoskedastic, inferences can be conducted using the usual OLS standard errors. Alternatively, without assuming homoskedasticity, heteroskedasticity-consistent  $t$ -statistics using standard errors can be derived using the approach of Halbert White (1980). The use of heteroskedasticity-consistent standard errors is advisable because there is no reason to expect the residuals of (26) to be homoskedastic.

Paul Asquith and David Mullins (1986) provide an example of this cross-sectional approach. The two day cumulative abnormal return for the announcement of an equity offering is regressed on the size of the offering as a percentage of the value of the total equity of the firm and on the cumulative abnormal re-

turn in the eleven months prior to the announcement month. They find that the magnitude of the (negative) abnormal return associated with the announcement of equity offerings is related to both these variables. Larger pre-event cumulative abnormal returns are associated with less negative abnormal returns and larger offerings are associated with more negative abnormal returns. These findings are consistent with theoretical predictions which they discuss.

Issues concerning the interpretation of the results can arise with the cross-sectional regression approach. In many situations, the event window abnormal return will be related to firm characteristics not only through the valuation effects of the event but also through a relation between the firm characteristics and the extent to which the event is anticipated. This can happen when investors rationally use the firm characteristics to forecast the likelihood of the event occurring. In these cases, a linear relation between the valuation effect of the event and the firm characteristic can be hidden. Paul Malatesta and Thompson (1985) and William Lanen and Thompson (1988) provide examples of this situation.

Technically, with the relation between the firm characteristics and the degree of anticipation of the event introduces a selection bias. The assumption that the regression residual is uncorrelated with the regressors breaks down and the OLS estimators are inconsistent. Consistent estimators can be derived by explicitly incorporating the selection bias. Sankarshan Acharya (1988) and B. Espen Eckbo, Vojislav Maksimovic, and Joseph Williams (1990) provide examples of this approach. N. R. Prabhala (1995) provides a good discussion of this problem and the possible solutions. He argues that, despite an incorrect specification, under weak conditions, the OLS ap-

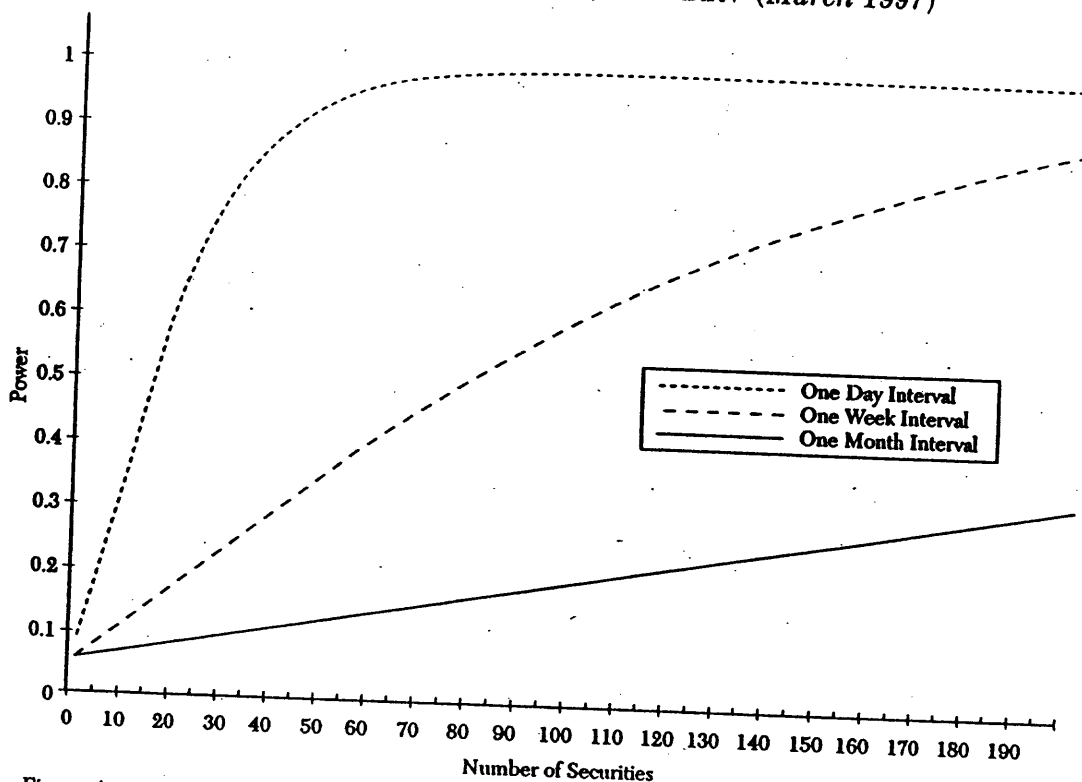


Figure 4. Power of event study test statistic  $\theta_1$  to reject the null hypothesis that the abnormal return is zero, for different sampling intervals, when the square root of the average variance of the abnormal return across firms is 4 percent for the daily interval. Size of test is 5 percent.

proach can be used for inferences and that the  $t$ -statistics can be interpreted as lower bounds on the true significance level of the estimates.

#### 10. Other Issues

A number of further issues often arise when conducting an event study. These issues include the role of the sampling interval, event date uncertainty, robustness, and some additional biases.

##### A. Role of Sampling Interval

Stock return data is available at different sampling intervals, with daily and monthly intervals being the most common. Given the availability of various intervals, the question of the gains of using

more frequent sampling arises. To address this question one needs to consider the power gains from shorter intervals. A comparison of daily versus monthly data is provided in Figure 4. The power of the test of no event effect is plotted against the alternative of an abnormal return of one percent for 1 to 200 securities. As one would expect given the analysis of Section 7, the decrease in power going from a daily interval to a monthly interval is severe. For example, with 50 securities the power for a 5 percent test using daily data is 0.94, whereas the power using weekly and monthly data is only 0.35 and 0.12 respectively. The clear message is that there is a substantial payoff in terms of increased power from reducing the sampling inter-



val. Dale Morse (1984) presents detailed analysis of the choice of daily versus monthly data and draws the same conclusion.

A sampling interval of one day is not the shortest interval possible. With the increased availability of transaction data, recent studies have used observation intervals of duration shorter than one day. However, the net benefit of intervals less than one day is unclear as some complications are introduced. Discussion of using transaction data for event studies is included in the work of Michael Barclay and Robert Litzenberger (1988).

#### B. *Inferences with Event-Date Uncertainty*

Thus far it is assumed that the event date can be identified with certainty. However, in some studies it may be difficult to identify the exact date. A common example is when collecting event dates from financial publications such as the *Wall Street Journal*. When the event announcement appears in the paper one can not be certain if the market was informed prior to the close of the market the prior trading day. If this is the case then the prior day is the event day, if not then the current day is the event day. The usual method of handling this problem is to expand the event window to two days—day 0 and day +1. While there is a cost to expanding the event window, the results in Section 6 indicated that the power properties of two day event windows are still good suggesting that the costs are worth bearing rather than to take the risk of missing the event.

Clifford Ball and Walter Torous (1988) have investigated the issue. They develop a maximum likelihood estimation procedure which accommodates event date uncertainty and examine results of their explicit procedure versus the informal procedure of expanding the event window. The results indicates that the

informal procedure works well and there is little to gain from the more elaborate estimation framework.

#### C. *Robustness*

The statistical analysis of Sections 4, 5, and 6 is based on assumption that returns are jointly normal and temporally independently and identically distributed. In this section, discussion of the robustness of the results to departures from this assumption is presented. The normality assumption is important for the exact finite sample results to hold. Without assuming normality, all results would be asymptotic. However, this is generally not a problem for event studies because for the test statistics, convergence to the asymptotic distributions is rather quick. Brown and Warner (1985) provide discussion of this issue.

#### D. *Other Possible Biases*

A number of possible biases can arise in the context of conducting an event study. Nonsynchronous trading can introduce a bias. The nontrading or nonsynchronous trading effect arises when prices are taken to be recorded at time intervals of one length when in fact they are recorded at time intervals of other possibly irregular lengths. For example, the daily prices of securities usually employed in event studies are generally "closing" prices, prices at which the last transaction in each of those securities occurred during the trading day. These closing prices generally do not occur at the same time each day, but by calling them "daily" prices, one is implicitly and incorrectly assuming that they are equally spaced at 24-hour intervals. This nontrading effect induces biases in the moments and co-moments of returns.

The influence of the nontrading effect on the variances and covariances of individual stocks and portfolios naturally feeds into a bias for the market model

beta. Myron Scholes and Williams (1977) present a consistent estimator of beta in the presence of nontrading based on the assumption that the true return process is uncorrelated through time. They also present some empirical evidence which shows the nontrading-adjusted beta estimates of thinly traded securities to be approximately 10 to 20 percent larger than the unadjusted estimates. However, for actively traded securities, the adjustments are generally small and unimportant.

Prem Jain (1986) considers the influence of thin trading on the distribution of the abnormal returns from the market model with the beta estimated using the Scholes-Williams approach. When comparing the distribution of these abnormal returns to the distribution of the abnormal returns using the usual OLS betas finds that the differences are minimal. This suggests that in general the adjustments for thin trading are not important.

The methodology used to compute the cumulative abnormal returns can induce an upward bias. The bias arises from the observation by observation rebalancing to equal weights implicit in the calculation of the aggregate cumulative abnormal return combined with the use of transaction prices which can represent both the bid and the offer side of the market. Marshall Blume and Robert Stambaugh (1983) analyze this bias and show that it can be important for studies using low market capitalization firms which have, in percentage terms, wide bid offer spreads. In these cases the bias can be eliminated by considering cumulative abnormal returns which represent buy and hold strategies.

### 11. Concluding Discussion

In closing, examples of event study successes and limitations are presented. Perhaps the most successful applications

have been in the area of corporate finance. Event studies dominate the empirical research in this area. Important examples include the wealth effects of mergers and acquisitions and the price effects of financing decisions by firms. Studies of these events typically focus on the abnormal return around the date of first announcement.

In the 1960s there was a paucity of empirical evidence on the wealth effects of mergers and acquisitions. For example, Henry Manne (1965) discusses the various arguments for and against mergers. At that time the debate centered on the extent to which mergers should be regulated in order to foster competition in the product markets. Manne argued that mergers represent a natural outcome in an efficiently operating market for corporate control and consequently provide protection for shareholders. He downplayed the importance of the argument that mergers reduce competition. At the conclusion of his article Manne suggested that the two competing hypotheses for mergers could be separated by studying the price effects of the involved corporations. He hypothesized that, if mergers created market power, one would observe price increases for both the target and acquirer. In contrast, if the merger represented the acquiring corporation paying for control of the target, one would observe a price increase for the target only and not for the acquirer. However, Manne concludes, in reference to the price effects of mergers, that "no data are presently available on this subject."

Since that time an enormous body of empirical evidence on mergers and acquisitions has developed which is dominated by the use of event studies. The general result is that, given a successful takeover, the abnormal returns of the targets are large and positive and the abnormal returns of the acquirer are close

to zero.  
docum  
turn f  
perce  
takeov  
the a  
close  
rell a  
mal r  
In the  
mal r  
Eckbo  
role  
plaini  
turns  
ing fi  
no ev  
comp  
he fi  
merg  
abno  
clude  
prod  
plana  
comp  
likely  
piric  
acqu  
chae  
and  
(198  
work  
A  
deve  
ing  
corp  
capi  
aver  
mag  
penc  
ing.  
a sa  
uity  
two  
perc  
the  
son  
two

to zero. Gregg Jarrell and Poulsen (1989) document that the average abnormal return for target shareholders exceeds 20 percent for a sample of 663 successful takeovers from 1960 to 1985. In contrast the abnormal returns for acquirers is close to zero. For the same sample, Jarrell and Poulsen find an average abnormal return of 1.14 percent for acquirers. In the 1980s they find the average abnormal return is negative at -1.10 percent. Eckbo (1983) explicitly addresses the role of increased market power in explaining merger related abnormal returns. He separates mergers of competing firms from other mergers and finds no evidence that the wealth effects for competing firms are different. Further, he finds no evidence that rivals of firms merging horizontally experience negative abnormal returns. From this he concludes that reduced competition in the product market is not an important explanation for merger gains. This leaves competition for corporate control a more likely explanation. Much additional empirical work in the area of mergers and acquisitions has been conducted. Michael Jensen and Richard Ruback (1983) and Jarrell, James Brickley, and Netter (1988) provide detailed surveys of this work.

A number of robust results have been developed from event studies of financing decisions by corporations. When a corporation announces that it will raise capital in external markets there is, on average, a negative abnormal return. The magnitude of the abnormal return depends on the source of external financing. Asquith and Mullins (1986) find for a sample of 266 firms announcing an equity issue in the period 1963 to 1981 the two day average abnormal return is -2.7 percent and on a sample of 80 firms for the period 1972 to 1982 Wayne Mikkelsen and Megan Partch (1986) find the two day average abnormal return is

-3.56 percent. In contrast, when firms decide to use straight debt financing, the average abnormal return is closer to zero. Mikkelsen and Partch (1986) find the average abnormal return for debt issues to be -0.23 percent for a sample of 171 issues. Findings such as these provide the fuel for the development of new theories. For example, in this case, the findings motivate the pecking order theory of capital structure developed by Stewart Myers and Nicholas Majluf (1984).

A major success related to those in the corporate finance area is the implicit acceptance of event study methodology by the U.S. Supreme Court for determining materiality in insider trading cases and for determining appropriate disgorgement amounts in cases of fraud. This implicit acceptance in the 1988 Basic, Incorporated v. Levinson case and its importance for securities law is discussed in Mitchell and Netter (1994).

There have also been less successful applications. An important characteristic of a successful event study is the ability to identify precisely the date of the event. In cases where the event date is difficult to identify or the event date is partially anticipated, studies have been less useful. For example, the wealth effects of regulatory changes for affected entities can be difficult to detect using event study methodology. The problem is that regulatory changes are often debated in the political arena over time and any accompanying wealth effects generally will gradually be incorporated into the value of a corporation as the probability of the change being adopted increases.

Larry Dann and Christopher James (1982) discuss this issue in the context of the impact of deposit interest rate ceilings for thrift institutions. In their study of changes in rate ceilings, they decide not to consider a change in 1973 because it was due to legislative action. Schipper

and Thompson (1983, 1985) also encounter this problem in a study of merger related regulations. They attempt to circumvent the problem of regulatory changes being anticipated by identifying dates when the probability of a regulatory change being passed changes. However, they find largely insignificant results leaving open the possibility the of absence of distinct event dates as the explanation of the lack of wealth effects.

Much has been learned from the body of research based on the use of event study methodology. In a general context, event studies have shown that, as would be expected in a rational marketplace, prices do respond to new information. As one moves forward, it is expected that event studies will continue to be a valuable and widely used tool in economics and finance.

## REFERENCES

- ACHARYA, SANKARSHAN. "A Generalized Econometric Model and Tests of a Signalling Hypothesis with Two Discrete Signals," *J. Finance*, June 1988, 43(2), pp. 413-29.
- ASHLEY, JOHN W. "Stock Prices and Changes in Earnings and Dividends: Some Empirical Results," *J. Polit. Econ.*, Feb. 1962, 70(1), pp. 82-85.
- ASQUITH, PAUL AND MULLINS, DAVID. "Equity Issues and Offering Dilution," *J. Finan. Econ.*, Jan./Feb. 1986, 15(1/2), pp. 61-89.
- BALL, CLIFFORD A. AND TOROUS, WALTER N. "Investigating Security-Price Performance in the Presence of Event-Date Uncertainty," *J. Finan. Econ.*, Oct. 1988, 22(1), pp. 123-53.
- BALL, RAY AND BROWN, PHILIP. "An Empirical Evaluation of Accounting Income Numbers," *J. Acc. Res.*, Autumn 1968, 6(2), pp. 159-78.
- BARCLAY, MICHAEL J. AND LITZENBERGER, ROBERT H. "Announcement Effects of New Equity Issues and the Use of Intraday Price Data," *J. Finan. Econ.*, May 1988, 21(1), pp. 71-99.
- BARKER, C. AUSTIN. "Effective Stock Splits," *Harvard Bus. Rev.*, Jan./Feb. 1956, 34(1), pp. 101-06.
- . "Stock Splits in a Bull Market," *Harvard Bus. Rev.*, May/June 1957, 35(3), pp. 72-79.
- . "Evaluation of Stock Dividends," *Harvard Bus. Rev.*, July/Aug. 1958, 36(4), pp. 99-114.
- BERNARD, VICTOR L. "Cross-Sectional Dependence and Problems in Inference in Market-Based Accounting Research," *J. Acc. Res.*, 1987, 25(1), pp. 1-48.
- BLUME, MARSHALL E. AND STAMBAUGH, ROBERT F. "Biases in Computed Returns: An Application to the Size Effect," *J. Finan. Econ.*, Nov. 1983, 12(3), pp. 387-404.
- BOEHMER, EKKEHART; MUSUMECI, JIM AND POULSEN, ANNETTE B. "Event-Study Methodology under Conditions of Event-Induced Variance," *J. Finan. Econ.*, Dec. 1991, 30(2), pp. 253-72.
- BROWN, STEPHEN J. AND WARNER, JEROLD B. "Measuring Security Price Performance," *J. Finan. Econ.*, Sept. 1980, 8(3), 205-58.
- . "Using Daily Stock Returns: The Case of Event Studies," *J. Finan. Econ.*, Mar. 1985, 14(1), pp. 3-31.
- BROWN, STEPHEN AND WEINSTEIN, MARK I. "Derived Factors in Event Studies," *J. Finan. Econ.*, Sept. 1985, 14(3), pp. 491-95.
- CAMPBELL, CYNTHIA J. AND WASLEY, CHARLES E. "Measuring Security Price Performance Using Daily NASDAQ Returns," *J. Finan. Econ.*, Feb. 1993, 33(1), pp. 73-92.
- COLLINS, DANIEL W. AND DENT, WARREN T. "A Comparison of Alternative Testing Methodologies Used In Capital Market Research," *J. Acc. Res.*, Spring 1984, 22(1), pp. 48-84.
- CORRADO, CHARLES. "A Nonparametric Test for Abnormal Security-Price Performance in Event Studies," *J. Finan. Econ.*, Aug. 1989, 23(2), pp. 385-95.
- DANN, LARRY Y. AND JAMES, CHRISTOPHER M. "An Analysis of the Impact of Deposit Rate Ceilings on the Market Values of Thrift Institutions," *J. Finance*, Dec. 1982, 37(5), pp. 1259-75.
- DOLLEY, JAMES CLAY. "Characteristics and Procedure of Common Stock Split-Ups," *Harvard Bus. Rev.*, Apr. 1933, 11, pp. 316-26.
- ECKBO, B. ESPEN. "Horizontal Mergers, Collusion, and Stockholder Wealth," *J. Finan. Econ.*, Apr. 1983, 11(1-4), pp. 241-73.
- ECKBO, B. ESPEN; MAKSIMOVIC, VOJISLAV AND WILLIAMS, JOSEPH. "Consistent Estimation of Cross-Sectional Models in Event Studies," *Rev. Financial Stud.*, 1990, 3(3), pp. 343-65.
- FAMA, EUGENE F. ET AL. "The Adjustment of Stock Prices to New Information," *Int. Econ. Rev.*, Feb. 1969, 10(1), pp. 1-21.
- FAMA, EUGENE F. AND FRENCH, KENNETH R. "Multifactor Explanations of Asset Pricing Anomalies," *J. Finance*, Mar. 1996, 51(1), pp. 55-84.
- JAIN, PREM. "Analyses of the Distribution of Security Market Model Prediction Errors for Daily Returns Data," *J. Acc. Res.*, Spring 1986, 24(1), pp. 76-96.
- JARRELL, GREGG A.; BRICKLEY, JAMES A. AND NETTER, JEFFRY M. "The Market for Corporate Control: The Empirical Evidence Since 1980," *J. Econ. Perspectives*, Winter 1988, 2(1), pp. 49-68.
- JARRELL, GREGG AND POULSEN, ANNETTE. "The Returns to Acquiring Firms in Tender Offers:



- Evidence from Three Decades," *Financial Management*, Autumn 1989, 18(3), pp. 12-19.
- JENSEN, MICHAEL C. AND RUBACK, RICHARD S. "The Market for Corporate Control: The Scientific Evidence," *J. Finan. Econ.*, Apr. 1983, 11(1-4), pp. 5-50.
- KANE, EDWARD J. AND UNAL, HALUK. "Change in Market Assessments of Deposit-Institution Riskiness," *J. Finan. Services Res.*, June 1988, 1(3), pp. 207-29.
- LANEN, WILLIAM N. AND THOMPSON, REX. "Stock Price Reactions as Surrogates for the Net Cash-Flow Effects of Corporate Policy Decisions," *J. Acc. Econ.*, Dec. 1988, 10(4), pp. 311-34.
- LINTNER, JOHN. "The Valuation of Risky Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Rev. Econ. Stat.*, Feb. 1965, 47(1), pp. 13-37.
- MACKINLAY, A. CRAIG. "On Multivariate Tests of the CAPM," *J. Finan. Econ.*, June 1987, 18(2), pp. 341-71.
- MALATESTA, PAUL H. AND THOMPSON, REX. "Partially Anticipated Events: A Model of Stock Price Reactions with an Application to Corporate Acquisitions," *J. Finan. Econ.*, June 1985, 14(2), pp. 237-50.
- MANNE, HENRY G. "Mergers and the Market for Corporate Control," *J. Polit. Econ.*, Apr. 1965, 73(2), pp. 110-20.
- MCQUEEN, GRANT AND ROLEY, VANCE. "Stock Prices, News, and Business Conditions," *Rev. Finan. Stud.*, 1993, 6(3), pp. 683-707.
- MIKKELSON, WAYNE H. AND PARTCH, MEGAN. "Valuation Effects of Security Offerings and the Issuance Process," *J. Finan. Econ.*, Jan./Feb. 1986, 15(1/2), pp. 31-60.
- MITCHELL, MARK L. AND NETTER, JEFFRY M. "The Role of Financial Economics in Securities Fraud Cases: Applications at the Securities and Exchange Commission," *Business Lawyer*, Feb. 1994, 49(2), pp. 545-90.
- MORSE, DALE. "An Econometric Analysis of the Choice of Daily Versus Monthly Returns In Tests of Information Content," *J. Acc. Res.*, Autumn 1984, 22(2), pp. 605-23.
- MYERS, JOHN H. AND BAKAY, ARCHIE J. "Influence of Stock Split-Ups on Market Price," *Harvard Bus. Rev.*, Mar. 1948, 26, pp. 251-55.
- MYERS, STEWART C. AND MAJLUF, NICHOLAS S. "Corporate Financing and Investment Decisions When Firms Have Information That Investors Do Not Have," *J. Finan. Econ.*, June 1984, 13(2), pp. 187-221.
- PATELL, JAMES M. "Corporate Forecasts of Earnings Per Share and Stock Price Behavior: Empirical Tests," *J. Acc. Res.*, Autumn 1976, 14(2), pp. 246-76.
- PRABHALA, N. R. "Conditional Methods in Event Studies and an Equilibrium Justification for Using Standard Event Study Procedures." Working Paper, Yale U., Sept. 1995.
- RITTER, JAY R. "Long-Run Performance of Initial Public Offerings," *J. Finance*, Mar. 1991, 46(1), pp. 3-27.
- ROSS, STEPHEN A. "The Arbitrage Theory of Capital Asset Pricing," *J. Econ. Theory*, Dec. 1976, 13(3), pp. 341-60.
- SCHIPPER, KATHERINE AND THOMPSON, REX. "The Impact of Merger-Related Regulations on the Shareholders of Acquiring Firms," *J. Acc. Res.*, Spring 1983, 21(1), pp. 184-221.
- . "The Impact of Merger-Related Regulations Using Exact Distributions of Test Statistics," *J. Acc. Res.*, Spring 1985, 23(1), pp. 408-15.
- SCHOLES, MYRON AND WILLIAMS, JOSEPH T. "Estimating Betas from Nonsynchronous Data," *J. Finan. Econ.*, Dec. 1977, 5(3), pp. 309-27.
- SCHWERT, G. WILLIAM. "Using Financial Data to Measure Effects of Regulation," *J. Law Econ.*, Apr. 1981, 24(1), pp. 121-58.
- SHARPE, WILLIAM F. "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *J. Finance*, Sept. 1964, 19(3), pp. 425-42.
- . *Portfolio theory and capital markets*. New York: McGraw-Hill, 1970.
- SHARPE, WILLIAM F.; ALEXANDER, GORDON J. AND BAILEY, JEFFERY V. *Investments*. Fifth Ed. Englewood Cliffs, NJ: Prentice-Hall, 1995.
- WHITE, HALBERT. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, 48(4), pp. 817-38.

# Exhibit 59

# The “File Drawer Problem” and Tolerance for Null Results

Robert Rosenthal  
Harvard University

For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the “file drawer problem” is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show non-significant results. Quantitative procedures for computing the tolerance for filed and future null results are reported and illustrated, and the implications are discussed.

Both behavioral researchers and statisticians have long suspected that the studies published in the behavioral sciences are a biased sample of the studies that are actually carried out (Bakan, 1967; McNemar, 1960; Smart, 1964; Sterling, 1959). The extreme view of this problem, the “file drawer problem,” is that the journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show nonsignificant (e.g.,  $p > .05$ ) results.

In the past there was very little one could do to assess the net effect of studies, tucked away in file drawers, that did not make the magic .05 level (Rosenthal & Gaito, 1963, 1964). Now, however, although no definitive solution to the problem is available, one can establish reasonable boundaries on the problem and estimate the degree of damage to any research conclusion that could be done by the file drawer problem.

This advance in our ability to cope with the file drawer is an outgrowth of the increasing interest of behavioral scientists in summarizing bodies of research literature sys-

tematically and quantitatively, both with respect to significance levels (Rosenthal, 1969, 1976, 1978) and with respect to effect-size estimation (Hall, 1978; Rosenthal, 1969, 1976; Rosenthal & Rosnow, 1975; Smith & Glass, 1977; Glass, Note 1). One hopes that this interest in summarizing entire research domains will lead to an improvement in book-keeping so that eventually all results will be recorded both with an estimate of effect size (e.g.,  $r$  or  $d$ ; Cohen, 1977) and with the level of significance obtained, or more practically, with the standard normal deviate ( $Z$ ) that corresponds to the obtained  $p$  (Rosenthal, 1978).<sup>1</sup> Future appraisals of research domains of the type found in *Psychological Bulletin* should give estimates of overall effect sizes and significance levels; these estimates of overall significance can provide a basis for coping with the file drawer problem.

## Tolerance for Future Null Results

Given any systematic quantitative review of the literature bearing on a particular hy-

Preparation of this article was supported in part by the National Science Foundation.

I would like to thank Judith A. Hall and Donald B. Rubin for their valuable improvements of an earlier version of this article.

Requests for reprints should be sent to Robert Rosenthal, Department of Psychology and Social Relations, Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138.

<sup>1</sup>Standard normal deviates ( $Z$ ) can be found by various methods, of which the following three are most often useful: (a) Obtain the exact  $p$  associated with the test statistic (e.g.,  $t$ ,  $F$ , or  $\chi^2$ ) and find the  $Z$  associated with that  $p$  in tables of the normal distribution; (b) if the effect size  $r$  or phi is given or can be computed,  $Z$  can be estimated by  $r(N)^{\frac{1}{2}}$ ; (c) if the effect size  $d$  is given or can be computed,  $Z$  can be estimated by  $[d^2/(d^2 + 4)]^{\frac{1}{2}}(N)^{\frac{1}{2}}$ .



pothesis, for example, that psychotherapy is effective (Glass, Note 1), that women are more sensitive than men to nonverbal cues (Hall, 1978), or that one person's expectation for another person's behavior can come to serve as self-fulfilling prophecy (Rosenthal, 1969, 1976), it is easy to calculate an overall probability, based on all the independent studies available to the reviewer, that the effect in question is "real," that is, not a Type I error (Rosenthal, 1978). The fundamental idea in coping with the file drawer problem is simply to calculate the number of studies averaging null results that must be in the file drawers before the overall probability of a Type I error is brought to any desired level of significance, say,  $p = .05$ . This number of filed studies, or the tolerance for future null results, is then evaluated for whether such a tolerance level is small enough to threaten the overall conclusion drawn by the reviewer. If the overall level of significance of the research review will be brought down to the level of *just significant* by the addition of just a few more null results, the finding is not resistant to the file drawer threat.

Computation

Perhaps the simplest, most useful way of computing the overall  $p$  of a set of research studies is the method of adding  $Z$ s (Cochran, 1954; Mosteller & Bush, 1954; Rosenthal, 1978). This method requires only that one add the standard normal deviates of  $Z$ s associated with the  $p$ s obtained and divide by the square root of the number of studies being combined. The result is itself a  $Z$  that can be entered in a table to find the associated overall  $p$ :

$$Z_c = k\bar{Z}_k/\sqrt{k} = \sqrt{k}\bar{Z}_k, \tag{1}$$

where  $Z_c$  is the new combined  $Z$ ,  $k$  is the number of studies combined, and  $\bar{Z}_k$  is the mean  $Z$  obtained for the  $k$  studies.

To find the number ( $X$ ) of new, filed, or unretrieved studies averaging null results required to bring the new overall  $p$  to any desired level, say, just significant at  $p = .05$

( $Z = 1.645$ ), one simply writes:

$$1.645 = k\bar{Z}_k/\sqrt{k + X}. \tag{2}$$

Rearrangement shows, then, that

$$X = (k/2.706)[k(\bar{Z}_k)^2 - 2.706]. \tag{3}$$

An alternative formula that may be more convenient when the sum of the  $Z$ s ( $\Sigma Z$ ) is given rather than the mean  $Z$  is as follows:  $X = [(\Sigma Z)^2 / 2.706] - k$ . One method based on counting rather than adding  $Z$ s may be easier to compute and can be employed when exact  $p$  levels are not available; but it is probably less powerful. If  $X$  is the number of new studies required to bring the overall  $p$  to .50 (not to .05),  $s$  is the number of summarized studies significant at  $p < .05$ , and  $n$  is the number of summarized studies not significant at .05, then  $X = 19s - n$ . Another conservative alternative when exact  $p$  levels are not available is to set  $Z = .00$  for any nonsignificant result and to set  $Z = 1.645$  for any result significant at  $p \leq .05$ .

Equations 1, 2, and 3 all assume that each of the  $k$  studies is independent of all other  $k - 1$  studies, at least in the sense of employing different sampling units. There are other senses of independence, however; for example, one can think of two or more studies conducted in a given laboratory as less independent than two or more studies conducted in different laboratories. Such non-independence can be assessed by intraclass correlations. Whether nonindependence of this type serves to increase Type I or Type II errors appears to depend in part on the relative magnitude of the  $Z$ s obtained from the studies that are correlated or too similar. If the correlated  $Z$ s are, on the average, as high (or higher) as the grand mean  $Z$  corrected for nonindependence, the combined  $Z$  one computes by treating all studies as independent will be too large. If the correlated  $Z$ s are, on the average, clearly low relative to the grand mean  $Z$  corrected for nonindependence, the combined  $Z$  one computes by treating all studies as independent will tend to be too small.

### Illustration

In 1969, 94 experiments examining the effects of interpersonal self-fulfilling prophecies were summarized (Rosenthal, 1969). The mean  $Z$  of these studies was 1.014,  $k$  was 94, and  $Z_c$  for the studies combined was  $9.83 = 94(1.014)/(94)^{1/2}$ .

How many new, filed, or unretrieved studies ( $X$ ) would be required to bring this very large  $Z$  down to a barely significant level ( $Z = 1.645$ )? By Equation 3,

$$X = (94/2.706) [94(1.014)^2 - 2.706] = 3,263.$$

One finds that 3,263 studies averaging null results ( $\bar{Z} = .00$ ) must be crammed into file drawers before one would conclude that the overall results were due to sampling bias in the studies summarized by the reviewer. In a more recent summary of the same area of research (Rosenthal, 1976), the mean  $Z$  of 311 studies was 1.180,  $k$  was 311, and  $X$  was 49,457! Thus, nearly 50,000 unreported studies averaging a null result would have to exist somewhere before the overall results could reasonably be ascribed to sampling bias.

### Discussion

There is both a sobering and a cheering lesson to be learned from careful study of Equation 3. The sobering lesson is that small numbers of studies that are not very significant, even when their combined  $p$  is significant, may well be misleading in that only a few studies filed away could change the combined significant result to a nonsignificant one. Thus, 15 studies averaging a  $Z$  of .50 have a combined  $p$  of .026; but if there were only 6 studies tucked away showing a mean  $Z$  of .00, the tolerance level for null results would be exceeded, and the significant result would become nonsignificant (i.e.,  $p > .05$ ). Or if there were 2 studies averaging a  $Z$  of 2.00, the combined  $p$  would be about .002; but uncovering 4 new studies averaging a  $Z$  of .00 would bring  $p$  into the *not significant* region.

The cheering lesson is that when the number of studies available grows large or the mean directional  $Z$  grows large, the file drawer hypothesis as a plausible rival hypothesis can be safely ruled out. If 300 studies are found to average a  $Z$  of +1.00, it would take 32,960 studies to bring the new combined  $p$  to a nonsignificant level; that many file drawers full is simply too improbable.

At the present time no firm guidelines can be given as to what constitutes an unlikely number of unretrieved or unpublished studies. For some areas of research 100 or even 500 unpublished and unretrieved studies may be a plausible state of affairs, whereas for others even 10 or 20 seems unlikely. Probably any rough and ready guide should be based partly on  $k$  so that as more studies are known it becomes more plausible that other studies in that area may be in those file drawers. Perhaps one could regard as resistant to the file drawer problem any combined results for which the tolerance level ( $X$ ) reaches  $5k + 10$ . This seems a conservative but reasonable tolerance level; the  $5k$  portion suggests that it is unlikely that the file drawers have more than five times as many studies as the reviewer, and the 10 sets the minimum number of studies that could be filed away at 15 (when  $k = 1$ ).

It appears that more and more reviewers of research literature are estimating average effect sizes and combined  $p$ s of the studies they summarize. It would be very helpful to readers if for each combined  $p$  they presented, reviewers also gave the tolerance for future null results associated with their overall significance level.

### Reference Note

1. Glass, G. V. *Primary, secondary, and meta-analysis of research*. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1976.

### References

- Bakan, D. *On method*. San Francisco: Jossey-Bass, 1967.
- Cochran, W. G. Some methods for strengthening the common  $\chi^2$  tests. *Biometrics*, 1954, 10, 417-451.

- Cohen, J. *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press, 1977.
- Hall, J. A. Gender effects in decoding nonverbal cues. *Psychological Bulletin*, 1978, *85*, 845-857.
- McNemar, Q. At random: Sense and nonsense. *American Psychologist*, 1960, *15*, 295-300.
- Mosteller, F. M., & Bush, R. R. Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology: Vol. 1. Theory and method*. Cambridge, Mass.: Addison-Wesley, 1954.
- Rosenthal, R. Interpersonal expectations. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research*. New York: Academic Press, 1969.
- Rosenthal, R. *Experimenter effects in behavioral research* (Enlarged ed.). New York: Irvington, 1976.
- Rosenthal, R. Combining results of independent studies. *Psychological Bulletin*, 1978, *85*, 185-193.
- Rosenthal, R., & Gaito, J. The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, 1963, *55*, 33-38.
- Rosenthal, R., & Gaito, J. Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports*, 1964, *15*, 570.
- Rosenthal, R., & Rosnow, R. L. *The volunteer subject*. New York: Wiley-Interscience, 1975.
- Smart, R. G. The importance of negative results in psychological research. *Canadian Psychologist*, 1964, *5*, 225-232.
- Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 1977, *32*, 752-760.
- Sterling, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, 1959, *54*, 30-34.

Received February 16, 1978 ■

#### Editorial Consultants for This Issue

- |                      |                       |                     |
|----------------------|-----------------------|---------------------|
| Mark I. Appelbaum    | John W. French        | Michael P. Maratsos |
| David Arenberg       | Paul A. Games         | Donald L. Meyer     |
| Pierce Barker        | Wendell R. Garner     | John Money          |
| Anthony Biglan       | Douglas R. Glasnapp   | Robert D. Nebes     |
| A. H. Black          | Goldine C. Gleser     | K. Daniel O'Leary   |
| R. Darrell Bock      | Harry F. Gollob       | Thomas Pettigrew    |
| Charles J. Brainerd  | Curtis Hardyck        | Peter Polson        |
| Jack W. Brehm        | Chester Harris        | Robert A. Rescorla  |
| Anthony Bryk         | Richard J. Harris     | Samuel H. Revusky   |
| Leonard S. Cahen     | John L. Horn          | Robert Rosenthal    |
| Angus Campbell       | Paul Horst            | John W. Schneider   |
| Russell M. Church    | Lawrence J. Hubert    | Barry Schwartz      |
| William V. Clemons   | Thomas J. Hummel      | Devendra Singh      |
| Gerald L. Clore      | Lloyd G. Humphreys    | Mary Lee Smith      |
| C. Keith Conners     | Douglas N. Jackson    | Brandt F. Steele    |
| James F. Crow        | Arthur R. Jensen      | John Thibaut        |
| Fred L. Damarin      | Anthony Kales         | Ross Traub          |
| Richard Darlington   | Gideon Keren          | William R. Uttal    |
| James H. Davis       | Walter Kintsch        | John P. Wanous      |
| Donald D. Dorfman    | Helena Chmura Kraemer | Paul H. Wender      |
| Alice H. Eagly       | C. C. Li              | Charles E. Werts    |
| Paul Ekman           | Joseph LoPiccolo      | Richard E. Whalen   |
| Jean-Claude Falmagne | R. Duncan Luce        | Jerry Wiggins       |
| N. T. Feather        | Michael Machover      | Rand Wilcox         |
| Joseph L. Fleiss     | Melvin Manis          | Herman A. Witkin    |
| Carl Frederiksen     |                       |                     |

# Exhibit 60

## **PDFlib PLOP: PDF Linearization, Optimization, Protection**

**Page inserted by evaluation version  
www.pdflib.com – sales@pdflib.com**

## EVENT-INDUCED VOLATILITY AND TESTS FOR ABNORMAL PERFORMANCE

Robert Savickas

*George Washington University*

### Abstract

I analyze a simple test statistic for mean abnormal returns in the presence of stochastic volatility during both event and nonevent windows and in the presence of event-induced variance increases. Unlike previous tests, the parametric test evaluated here does not require that the volatility effect of the event be the same across all securities. Simulations show that the test exhibits nontrivial gains in power over previously developed parametric and nonparametric tests, and the true null hypothesis is rejected at appropriate levels.

*JEL Classifications:* G14, C10

### I. Introduction

Since the publication of Fama et al. (1969), the event-study method with its many extensions has become an important tool in finance, accounting, and economics. Binder (1998), MacKinlay (1997), and Peterson (1989) provide detailed reviews of various modifications to the original method.

A reoccurring motif in the literature is stochastic volatility around the event period. Brown and Warner (1980, 1985) point out that increases in variance may result in misspecification of the traditional test statistics and that the power of tests can be improved by appropriately modeling the volatility process. In the approach of Boehmer, Musumeci, and Poulsen (1991), the event-period returns are standardized by the estimation-period standard deviation, and the cross-sectional mean of the standardized returns is divided by their cross-sectional standard deviation to yield the test statistic. This approach implicitly assumes that the event-induced variance is the same for all securities in the sample. Corrado (1989) proposes a nonparametric test to accommodate event-induced variance. Simulations in articles both by Boehmer, Musumeci, and Poulsen and by Corrado demonstrate higher power of these tests relative to the traditional approach.

---

I wish to thank Shane Corwin, Jimmy Hilliard, Jeff Netter, Annette Poulsen, the anonymous referee, and especially the editor, William T. Moore, for helpful suggestions on the earlier drafts. All remaining errors are my own.

The generalized autoregressive conditional heteroskedasticity (GARCH) of Bollerslev (1986) has become an important tool in financial analyses. As a result, Brockett, Chen, and Garven (1999) develop an event-study method that assumes a market model with GARCH effects and time-varying slope ( $\beta$ ). They, however, ignore the importance of event-induced variance, a phenomenon that is emphasized in Brown and Warner (1980, 1985), Boehmer, Musumeci, and Poulsen (1991), and Corrado (1989).

In the present article, both the conditionally heteroskedastic behavior of volatility and the event-induced variance increase are addressed in a single model. I use a GARCH(1,1) model with dummy variables to evaluate a simple test statistic that accounts for the stochastic behavior of volatility during both event and nonevent periods. The test does not require the volatility effect to be the same across firms in the sample. The test is easy to implement but has substantially higher rejection rates of a false null hypothesis than do the previous tests.

## II. Test Statistics

In this section I begin by summarizing three tests statistics that are frequently used in event studies: (1) the traditional testing technique; (2) the standardized cross-sectional approach, as in Boehmer, Musumeci, and Poulsen (1991); and (3) the mean rank approach, as in Corrado (1989). The GARCH-based method for incorporating the event-induced variance is introduced in the last part of the section.

In the following discussion, the abnormal return  $A_{i,t}$  for security  $i$  at day  $t$  is obtained by fitting a benchmark model—for example, the market model, the mean-adjusted returns model, or the market-adjusted returns model—to the returns series.  $T$  is the number of daily observations in the estimation period, and  $N$  is the number of securities in the sample.

### *Traditional Approach*

The cross-sectional average of abnormal returns is divided by the time-series estimate of its standard deviation as in Brown and Warner (1980) to obtain the  $t$ -distributed test statistic:

$$\text{test}_1 = \sum_{i=1}^N \frac{A_{i,t}}{N} \bigg/ \sqrt{\frac{1}{N^2(T-1)} \cdot \sum_{i=1}^N \sum_{\tau=1}^T \left( A_{i,\tau} - \sum_{j=1}^T A_{i,j} / T \right)^2}. \quad (1)$$

The statistic is distributed Student- $t$  with  $T-1$  degrees of freedom.



*Standardized Cross-Sectional Approach*

The abnormal returns  $A_{i,t}$  are standardized by the time-series estimate of the standard deviation for each security  $i$  to obtain the standardized residual  $S_{i,t}$ . The cross-sectional mean of  $S_{i,t}$  is divided by the cross-sectional estimate of its standard deviation as in Boehmer, Musumeci, and Poulsen (1991) to obtain the following test statistic:

$$\text{test}_2 = \sum_{i=1}^N \frac{S_{i,t}}{N} \Bigg/ \sqrt{\frac{1}{N(N-1)} \cdot \sum_{i=1}^N \left( S_{i,t} - \sum_{j=1}^N S_{j,t}/N \right)^2}, \quad (2)$$

where

$$S_{i,t} = A_{i,t} \Bigg/ \sqrt{\frac{1}{(T-1)} \cdot \sum_{\tau=1}^T \left( A_{i,\tau} - \sum_{j=1}^T A_{i,j}/T \right)^2}. \quad (3)$$

The statistic in equation (2) is distributed Student- $t$  with  $N-1$  degrees of freedom. It is common to then divide this test statistic by a term that adjusts for forecast errors.

*Mean Rank Approach*

The mean rank approach uses the entire time-series sample consisting of  $Q$  observations, without splitting it into the estimation period (with  $T$  observations) and the event period (with  $Q-T$  observations). All abnormal returns for each security are ranked based on their relative magnitudes. The smallest  $A_{i,t}$  is assigned a rank of 1:  $K_{i,t} = 1$ . If the abnormal return at time  $t$  is greater than the abnormal return at time  $j$ ,  $A_{i,t} > A_{i,j}$ , its rank will also be higher,  $K_{i,t} > K_{i,j}$ . If two abnormal returns are equal, they receive the same rank. The deviation of the cross-sectional estimate of the average rank from the theoretical average rank (under the null hypothesis, the theoretical average rank is equal to  $(Q+1)/2$ ) is divided by the time-series estimate of its standard deviation as in Corrado (1989) to obtain the test statistic:

$$\text{test}_3 = \left( \sum_{i=1}^N \frac{K_{i,t}}{N} - \frac{Q+1}{2} \right) \Bigg/ \sqrt{\frac{1}{Q} \sum_{\tau=1}^Q \left( \sum_{i=1}^N \frac{K_{i,\tau}}{N} - \frac{Q+1}{2} \right)^2}. \quad (4)$$

This statistic is standard normal.

*GARCH-Based Approach*

The GARCH-based approach involves estimating the following model using the time series of returns  $R_{i,t}$  and  $R_{m,t}$  for security  $i$  and market index  $m$ , respectively:

$$R_{i,t} = \alpha_i + \beta_i \cdot R_{m,t} + \gamma_i \cdot D_t + \eta_{i,t}, \quad \eta_{i,t} | \Omega_t \sim N(0, h_{i,t}), \quad (5)$$

$$h_{i,t} = a_i + b_i \cdot h_{i,t-1} + c_i \cdot \eta_{i,t-1}^2 + d_i \cdot D_t, \quad (6)$$

where  $\alpha_i$ ,  $\beta_i$ ,  $\gamma_i$ ,  $a_i$ ,  $b_i$ ,  $c_i$ , and  $d_i$  are parameters to be estimated;  $D_t$  is an indicator variable that equals 1 if  $t$  is an event day, and 0 otherwise; and  $\Omega_t$  is the set of information available at time  $t$ . The information set  $\Omega_t$  consists of all information available at time  $t$ , including all the current and previous market and security returns  $R_{m,u}$  and  $R_{i,u}$  for all  $u \leq t$ , current and previous volatility estimates  $h_{i,u}$  for all  $u \leq t$ , and current and previous error estimates  $\eta_{i,u}$  for all  $u \leq t$ . Note that the model can easily accommodate several indicator variables, each for different event periods (e.g.,  $D_{t,1}$  for period  $[-10, -3]$ ,  $D_{t,2}$  for period  $[-2, 2]$ ,  $D_{t,3}$  for period  $[3, 10]$ ). As in test<sub>3</sub>, the entire data set is used for model estimation.

In this framework, the mean of the market model residual  $R_{i,t} - \alpha_i - \beta_i R_{m,t}$  during the event period will be reflected in the estimate of  $\gamma_i$ , because by construction the mean of the disturbance  $\eta_{i,t}$  must be zero. If the null hypothesis of zero mean abnormal return is true, the estimate of  $\gamma_i$  must be close to zero. The precise meaning of “close” depends on the volatility of the market model residual reflected in  $h_{i,t}$ , which now incorporates the event-induced variance through the coefficient  $d_i$ . Dividing the estimated mean  $\hat{\gamma}_i$  of abnormal return for each security by the estimated standard deviation of abnormal return (not the standard deviation of the estimate)  $\hat{h}_i$  will result in a metric that accounts for security-specific event-induced volatility as well as return. The cross-sectional test statistic can now be formulated analogously to test<sub>2</sub>:

$$\text{test}_4 = \sum_{i=1}^N \frac{S_{i,t}}{N} \bigg/ \sqrt{\frac{1}{N(N-1)} \cdot \sum_{i=1}^N \left( S_{i,t} - \sum_{j=1}^N S_{j,t}/N \right)^2}, \quad (7)$$

where

$$S_{i,t} = \frac{\hat{\gamma}_i}{\sqrt{\hat{h}_{i,t}}}. \quad (8)$$

The test statistic in equation (7) is Student- $t$  distributed with  $N-1$  degrees of freedom. This approach allows the event-induced volatility effect to be different across all securities because  $d_i \neq d_j$  for  $i \neq j$ . Furthermore, unlike in the other tests considered, each security's variance is allowed to be stochastic outside the event period. These factors contribute to the potential gain in test power.

### III. Simulations

The Center for Research in Security Prices (CRSP) Daily Returns File is used to generate 250 portfolios of 50 randomly (with replacement) sampled

TABLE 1. Rejection Rates of Test Statistics: No Event-Induced Returns, No Event-Induced Variance.

Event Day	Significance Level		
	1%	5%	10%
Panel A. Test <sub>1</sub> : Traditional			
Day 1	4.00%	7.60%	9.60%
Day 2	4.00	8.80	13.20
Day 3	2.00	5.60	13.20
Day 4	4.40	8.40	12.00
Day 5	2.00	4.40	11.60
Panel B. Test <sub>2</sub> : Standardized Cross-Sectional			
Day 1	0.40	3.60	8.40
Day 2	0.80	4.80	9.60
Day 3	0.80	6.00	12.40
Day 4	1.20	4.00	8.40
Day 5	0.00	3.20	8.80
Panel C. Test <sub>3</sub> : Mean Rank			
Day 1	0.40	5.20	10.00
Day 2	0.80	3.20	8.40
Day 3	1.60	6.00	13.20
Day 4	1.20	6.00	10.80
Day 5	0.00	2.80	10.00
Panel D. Test <sub>4</sub> : GARCH-Based			
Day 1	0.80	4.40	11.20
Day 2	0.40	4.40	10.00
Day 3	0.40	4.40	10.00
Day 4	1.20	4.00	9.60
Day 5	1.20	4.80	10.00

Note: This table shows the rejection rates for four test statistics when an event causes no increase in returns and no increase in variance. Test<sub>1</sub> is as in Brown and Warner (1980), test<sub>2</sub> is as in Boehmer, Musumeci, and Poulsen (1991), test<sub>3</sub> is as in Corrado (1989), and test<sub>4</sub> is studied in this article. The rejection rates are computed for each of the five event days for three significance levels. These results are based on a 250-portfolio sample of 50 randomly sampled securities each; a 250-day sample is used for each security.

securities each. The first day of the time-series sample is randomly chosen for each security. A time series of 250 geometric returns for each security and the CRSP equally weighted index<sup>1</sup> is computed from the arithmetic returns. The sampling process excludes securities with missing returns during the 250-day interval.<sup>2</sup>

<sup>1</sup>The conclusions are not changed when the value-weighted and the S&P 500 indexes are used.  
<sup>2</sup>Note that the GARCH(1,1) process can be estimated by dropping the missing observations for each security, but this may affect the consistency of the estimates.

TABLE 2. Rejection Rates of Test Statistics: Event-Induced Returns, but No Event-Induced Variance.

Event Day	Significance Level		
	1%	5%	10%
Panel A. Test <sub>1</sub> : Traditional			
Day 1	16.40%	30.80%	42.80%
Day 2	10.40	29.60	41.20
Day 3	15.20	28.80	40.40
Day 4	17.20	35.20	49.20
Day 5	16.00	34.40	44.00
Panel B. Test <sub>2</sub> : Standardized Cross-Sectional			
Day 1	42.80	70.80	81.60
Day 2	47.20	73.60	80.80
Day 3	46.00	68.80	78.00
Day 4	46.00	69.60	80.00
Day 5	44.80	66.00	76.80
Panel C. Test <sub>3</sub> : Mean Rank			
Day 1	88.00	94.40	98.00
Day 2	87.20	95.20	98.00
Day 3	82.00	94.00	98.00
Day 4	88.00	98.00	98.80
Day 5	83.60	94.80	96.40
Panel D. Test <sub>4</sub> : GARCH-Based			
Day 1	99.20	100.00	100.00
Day 2	99.60	100.00	100.00
Day 3	99.60	100.00	100.00
Day 4	100.00	100.00	100.00
Day 5	100.00	100.00	100.00

Note: This table shows the rejection rates for four test statistics when an event causes an increase in returns and no increase in variance. Test<sub>1</sub> is as in Brown and Warner (1980), test<sub>2</sub> is as in Boehmer, Musumeci, and Poulsen (1991), test<sub>3</sub> is as in Corrado (1989), and test<sub>4</sub> is studied in this article. The rejection rates are computed for each of the five event days for three significance levels. These results are based on a 250-portfolio sample of 50 randomly sampled securities each; a 250-day sample is used for each security.

The last five observations in the entire 250-day sample are designated as the event period, during which the security-specific event-induced volatility and returns are introduced for each day individually. The abnormal volatility for security  $i$  on day  $t$  is introduced by multiplying the return  $R_{i,t}$  by the square root of a uniform random number in the interval  $[0.75AV_i, 1.25AV_i]$ , where  $AV_i$  is a realization of a uniform random number in the range  $[1.5, 2]$ . Thus, the variable  $AV_i$  represents the average multiplicative event-induced variance increase for security  $i$ . Each day's multiplicative variance increase for security  $i$  varies between 75% and 125% of the average.

TABLE 3. Rejection Rates of Test Statistics: Event-Induced Variance, but No Event-Induced Returns.

Event Day	Significance Level		
	1%	5%	10%
Panel A. Test <sub>1</sub> : Traditional			
Day 1	7.20%	16.40%	22.80%
Day 2	5.60	13.60	22.00
Day 3	10.00	20.40	27.60
Day 4	5.60	13.60	18.80
Day 5	7.20	14.40	22.40
Panel B. Test <sub>2</sub> : Standardized Cross-Sectional			
Day 1	0.40	4.40	8.40
Day 2	0.40	3.20	8.80
Day 3	1.60	6.40	11.60
Day 4	0.00	2.80	7.60
Day 5	0.00	3.60	9.20
Panel C. Test <sub>3</sub> : Mean Rank			
Day 1	1.60	6.00	11.20
Day 2	1.60	6.40	14.00
Day 3	2.00	8.00	13.20
Day 4	0.80	4.40	7.60
Day 5	1.60	5.20	10.00
Panel D. Test <sub>4</sub> : GARCH-Based			
Day 1	0.80	5.60	12.40
Day 2	0.80	5.60	11.20
Day 3	0.80	6.00	10.00
Day 4	0.80	5.60	11.60
Day 5	0.40	5.20	9.60

Note: This table shows the rejection rates for four test statistics when an event causes an increase in variance and no increase in returns. Test<sub>1</sub> is as in Brown and Warner (1980), test<sub>2</sub> is as in Boehmer, Musumeci, and Poulsen (1991), test<sub>3</sub> is as in Corrado (1989), and test<sub>4</sub> is studied in this article. The rejection rates are computed for each of the five event days for three significance levels. These results are based on a 250-portfolio sample of 50 randomly sampled securities each; a 250-day sample is used for each security.

When abnormal volatility is introduced, each abnormal return is simulated by adding a uniform random number in the interval  $[0.75AV_i, 1.25AV_i]$ , where  $AR_i$  is a realization of a uniform random number in the range  $[0.5\%, 1.5\%]$ , to the return  $R_{i,t}$  that is already multiplied by the square root of the variance increase.

This procedure assures that each event day for each security will have a different event-induced return and event-induced volatility. Other ranges for the random number generator and different values for event-induced returns and variance do not alter the conclusions of the study.

TABLE 4. Rejection Rates of Test Statistics: Event-Induced Returns and Event-Induced Variance.

Event Day	Significance Level		
	1%	5%	10%
Panel A. Test <sub>1</sub> : Traditional			
Day 1	22.00%	35.60%	42.40%
Day 2	19.20	36.40	45.60
Day 3	20.80	40.80	49.20
Day 4	28.00	40.80	48.80
Day 5	18.80	33.20	42.00
Panel B. Test <sub>2</sub> : Standardized Cross-Sectional			
Day 1	24.00	48.40	63.20
Day 2	21.20	42.40	56.80
Day 3	20.40	47.60	59.60
Day 4	28.00	49.60	57.20
Day 5	25.20	45.60	54.80
Panel C. Test <sub>3</sub> : Mean Rank			
Day 1	72.40	89.20	91.60
Day 2	72.80	86.80	94.00
Day 3	74.80	90.40	94.00
Day 4	71.20	88.00	96.00
Day 5	68.80	86.40	92.00
Panel D. Test <sub>4</sub> : GARCH-Based			
Day 1	98.00	99.20	99.20
Day 2	98.00	99.60	99.60
Day 3	98.40	99.60	99.60
Day 4	98.40	99.60	99.60
Day 5	98.40	99.60	99.60

Note: This table shows the rejection rates for four test statistics when an event causes both an increase in returns and an increase in variance. Test<sub>1</sub> is as in Brown and Warner (1980), test<sub>2</sub> is as in Boehmer, Musumeci, and Poulsen (1991), test<sub>3</sub> is as in Corrado (1989), and test<sub>4</sub> is studied in this article. The rejection rates are computed for each of the five event days for three significance levels. These results are based on a 250-portfolio sample of 50 randomly sampled securities each; a 250-day sample is used for each security.

IV. Empirical Results

The four tests are analyzed under four scenarios: (1) no event-induced excess returns and no variance increase, (2) event-induced returns but no variance, (3) event-induced variance but no returns, and (4) both event-induced returns and variance. The null hypothesis of zero mean abnormal returns is rejected by a particular test if the test’s two-tailed *p*-value is below a given significance level. A rejection rate is computed as the percentage of all portfolios for which a test rejects the null hypothesis. Tables 1–4 present the results.

TABLE 5. Rejection Rates of Test Statistics: No Event-Induced Returns, No Event-Induced Variance: Nasdaq Sample.

Event Day	Significance Level		
	1%	5%	10%
Panel A. Test <sub>1</sub> : Traditional			
Day 1	1.20%	4.80%	10.80%
Day 2	2.80	7.20	14.00
Day 3	4.00	9.20	13.20
Day 4	3.20	8.40	15.60
Day 5	2.00	7.20	11.60
Panel B. Test <sub>2</sub> : Standardized Cross-Sectional			
Day 1	1.60	2.80	8.40
Day 2	0.40	4.40	12.80
Day 3	0.80	2.40	9.60
Day 4	0.80	5.20	11.20
Day 5	0.80	2.80	8.00
Panel C. Test <sub>3</sub> : Mean Rank			
Day 1	0.40	3.60	8.80
Day 2	0.80	4.00	8.40
Day 3	0.40	4.80	6.80
Day 4	0.80	4.40	9.20
Day 5	0.40	4.40	7.20
Panel D. Test <sub>4</sub> : GARCH-Based			
Day 1	1.20	6.00	10.80
Day 2	1.20	5.60	10.40
Day 3	1.20	5.60	10.80
Day 4	0.80	5.60	11.20
Day 5	1.20	6.00	12.00

Note: This table shows the rejection rates for four test statistics when an event causes no increase in returns and no increase in variance. Test<sub>1</sub> is as in Brown and Warner (1980), test<sub>2</sub> is as in Boehmer, Musumeci, and Poulsen (1991), test<sub>3</sub> is as in Corrado (1989), and test<sub>4</sub> is studied in this article. The rejection rates are computed for each of the five event days for three significance levels. These results are based on a 250-portfolio sample of 50 randomly sampled Nasdaq securities each; a 250-day sample is used for each security.

Each of the tables consists of four panels. A panel displays the rejection rates for a particular test at three conventional significance levels for each of the five event days. Table 1 shows that when no event-induced returns or volatility are present, all four tests reject the true null hypothesis at approximately correct levels, equal to the size of the test.

When an event causes a change in abnormal returns but not volatility (Table 2), all four tests reject the null hypothesis at high levels. As documented in the previous literature (e.g., see Boehmer, Musumeci, and Poulsen 1991;



TABLE 6. Rejection Rates of Test Statistics: Event-Induced Returns, but No Event-Induced Variance: Nasdaq Sample.

Event Day	Significance Level		
	1%	5%	10%
Panel A. Test <sub>1</sub> : Traditional			
Day 1	18.00%	32.80%	44.40%
Day 2	14.80	30.00	41.20
Day 3	15.20	30.80	40.80
Day 4	16.80	33.60	43.20
Day 5	12.40	30.00	42.00
Panel B. Test <sub>2</sub> : Standardized Cross-Sectional			
Day 1	30.80	58.40	66.40
Day 2	29.20	55.60	70.00
Day 3	31.60	56.40	66.40
Day 4	32.40	54.80	68.00
Day 5	32.00	53.20	66.40
Panel C. Test <sub>3</sub> : Mean Rank			
Day 1	84.40	97.20	98.80
Day 2	85.60	96.40	98.80
Day 3	85.20	94.80	97.60
Day 4	83.60	93.20	96.40
Day 5	82.80	94.80	98.00
Panel D. Test <sub>4</sub> : GARCH-Based			
Day 1	97.20	99.60	99.60
Day 2	97.60	99.20	99.20
Day 3	98.80	99.60	99.60
Day 4	98.40	99.20	99.20
Day 5	98.40	99.60	99.60

Note: This table shows the rejection rates for four test statistics when an event causes an increase in returns and no increase in variance. Test<sub>1</sub> is as in Brown and Warner (1980), test<sub>2</sub> is as in Boehmer, Musumeci, and Poulsen (1991), test<sub>3</sub> is as in Corrado (1989), and test<sub>4</sub> is studied in this article. The rejection rates are computed for each of the five event days for three significance levels. These results are based on a 250-portfolio sample of 50 randomly sampled Nasdaq securities each; a 250-day sample is used for each security.

Corrado 1989), test<sub>2</sub> and test<sub>3</sub> outperform test<sub>1</sub>. The test studied here, test<sub>4</sub>, exhibits significantly greater rejection rates than the other three tests.

When an event causes abnormal volatility but not abnormal returns (Table 3), the classical misspecification of the traditional test becomes evident. The traditional test, test<sub>1</sub>, overrejects the true null hypothesis of zero mean abnormal returns, whereas the other three tests are immune to misspecification.

Finally, when an event causes both the returns and the volatilities to increase (Table 4), all tests reject the false null hypothesis at high levels. The ranking of tests based on their rejection rates is the same as in Table 2, with the GARCH-based

**TABLE 7. Rejection Rates of Test Statistics: Event-Induced Variance, but No Event-Induced Returns: Nasdaq Sample.**

Event Day	Significance Level		
	1%	5%	10%
Panel A. Test <sub>1</sub> : Traditional			
Day 1	8.40%	15.60%	22.00%
Day 2	7.20	20.00	25.60
Day 3	6.80	15.60	24.40
Day 4	8.80	22.40	30.80
Day 5	4.80	13.60	23.60
Panel B. Test <sub>2</sub> : Standardized Cross-Sectional			
Day 1	1.20	6.00	12.00
Day 2	0.40	2.00	8.40
Day 3	1.60	4.80	8.40
Day 4	0.80	6.40	13.20
Day 5	0.40	4.40	6.80
Panel C. Test <sub>3</sub> : Mean Rank			
Day 1	1.20	8.80	14.80
Day 2	1.20	5.60	12.00
Day 3	2.40	7.20	14.00
Day 4	1.60	7.20	13.60
Day 5	0.00	5.20	9.20
Panel D. Test <sub>4</sub> : GARCH-Based			
Day 1	0.80	6.00	12.40
Day 2	0.80	5.60	11.60
Day 3	0.80	6.00	12.40
Day 4	0.40	6.40	11.60
Day 5	0.40	6.40	11.60

Note: This table shows the rejection rates for four test statistics when an event causes an increase in variance and no increase in returns. Test<sub>1</sub> is as in Brown and Warner (1980), test<sub>2</sub> is as in Boehmer, Musumeci, and Poulsen (1991), test<sub>3</sub> is as in Corrado (1989), and test<sub>4</sub> is studied in this article. The rejection rates are computed for each of the five event days for three significance levels. These results are based on a 250-portfolio sample of 50 randomly sampled Nasdaq securities each; a 250-day sample is used for each security.

test<sub>4</sub> outperforming the other three tests. For example, the rejection rates of the 1% GARCH-based test are 24% to 30% higher than those of the 1% mean rank test, 70% to 78% higher than those of the 1% standardized cross-sectional test, and 70% to 80% higher than those of the traditional test.

**V. Simulations with Nasdaq Data**

A detailed discussion of structural differences between auction (New York Stock Exchange and American Stock Exchange) and dealer (Nasdaq) markets is

TABLE 8. Rejection Rates of Test Statistics: Event-Induced Returns and Event-Induced Variance: Nasdaq Sample.

Event Day	Significance Level		
	1%	5%	10%
Panel A. Test <sub>1</sub> : Traditional			
Day 1	17.20%	31.20%	43.20%
Day 2	18.80	31.20	40.80
Day 3	18.00	32.80	41.60
Day 4	16.40	29.20	39.60
Day 5	18.40	36.00	46.80
Panel B. Test <sub>2</sub> : Standardized Cross-Sectional			
Day 1	13.20	36.40	48.40
Day 2	14.80	33.20	47.60
Day 3	16.00	38.00	50.00
Day 4	14.00	33.20	46.40
Day 5	22.40	43.60	54.40
Panel C. Test <sub>3</sub> : Mean Rank			
Day 1	80.40	91.60	96.80
Day 2	80.40	94.00	96.00
Day 3	76.00	89.60	94.00
Day 4	75.20	92.40	94.40
Day 5	78.40	92.40	96.40
Panel D. Test <sub>4</sub> : GARCH-Based			
Day 1	92.40	98.40	98.80
Day 2	94.00	98.40	98.80
Day 3	94.40	98.40	98.80
Day 4	94.80	98.80	98.80
Day 5	94.40	98.40	99.20

Note: This table shows the rejection rates for four test statistics when an event causes both an increase in returns and an increase in variance. test<sub>1</sub> is as in Brown and Warner (1980), test<sub>2</sub> is as in Boehmer, Musumeci, and Poulsen (1991), test<sub>3</sub> is as in Corrado (1989), and test<sub>4</sub> is studied in this article. The rejection rates are computed for each of the five event days for three significance levels. These results are based on a 250-portfolio sample of 50 randomly sampled Nasdaq securities each; a 250-day sample is used for each security.

beyond the scope of this article, but these differences may affect the statistical properties of daily securities returns. Campbell and Wasley (1993) document substantial departures from normality in daily Nasdaq returns for both individual securities and portfolios. They compare the portfolio and standardized tests from Brown and Warner (1985) with the Corrado (1989) test and find that the first two tests are misspecified for Nasdaq data, whereas the Corrado nonparametric test is not. Campbell and Wasley recommend the use of the latter test for Nasdaq data.

Although the conditional distribution of next day's returns is normal under the GARCH process, the unconditional distribution is nonnormal. Because the

extensive GARCH literature indicates a good fit of the model to actual data (e.g., see Akgiray 1989), the GARCH-based test statistic for abnormal returns may be well specified for the Nasdaq data. In this section, simulations similar to those in the previous section are performed. The sample includes only Nasdaq securities, and the market index returns are those for the Nasdaq Composite Index provided on CRSP.

Tables 5 to 8 present the results. The conclusions drawn from the Nasdaq-based simulations are analogous to those in the previous section. The traditional test is misspecified in the presence of event-induced volatility, whereas the other three tests are not. Again, the GARCH-based technique outperforms the other three approaches in terms of test power.

## VI. Conclusion

The performance of four tests of abnormal returns during events is studied. The four tests are: (1) the traditional (Brown and Warner 1980), (2) the standardized cross-sectional (Boehmer, Musumeci, and Poulsen 1991), (3) the mean rank (Corrado 1989), and (4) the GARCH-based approach. Consistent with the previous literature, the traditional test is misspecified in the presence of event-induced variance. The mean rank test derives its power from the absence of a specific distributional assumption, producing higher rejection rates than those of the standardized cross-sectional test. The GARCH-based approach studied here explicitly models the volatility process and event-induced variance increases, which allows it to have the highest test power with the appropriate levels of Type I error. This approach rejects the false null hypothesis significantly more frequently than do the pre-existing tests whereas the rejection rates of a true null are close to the size of the test. Finally, the GARCH-based test remains well specified with the Nasdaq data.

## References

- Akgiray, V., 1989, Conditional heteroscedasticity in time series of stock returns: Evidence and forecasts, *Journal of Business* 62, 55–80.
- Binder, J., 1998, The event study methodology since 1969, *Review of Quantitative Finance and Accounting* 11, 111–37.
- Boehmer, E., J. Musumeci, and A. Poulsen, 1991, Event-study methodology under conditions of event-induced variance, *Journal of Financial Economics* 30, 253–72.
- Bollerslev, T., 1986, Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* 31, 307–27.
- Brockett, P., H. Chen, and J. Garven, 1999, A new stochastically flexible event methodology with application to Proposition 103, *Insurance, Mathematics and Economics* 25, 197–217.
- Brown, S. and J. Warner, 1980, Measuring security price performance, *Journal of Financial Economics* 8, 205–58.
- , 1985, Using daily stock returns. The case of event studies. *Journal of Financial Economics* 14, 3–31.

- Campbell, C. and C. Wasley, 1993, Measuring security price performance using daily Nasdaq returns, *Journal of Financial Economics* 33, 73–92.
- Corrado, C., 1989, A nonparametric test for abnormal security-price performance in event studies, *Journal of Financial Economics* 23, 385–95.
- Fama, E., L. Fisher, M. Jensen, and R. Roll, 1969, The adjustment of stock prices to new information, *International Economic Review* 10, 1–21.
- MacKinlay, A., 1997, Event studies in economics and finance, *Journal of Economic Literature* 35, 13–39.
- Peterson, P., 1989, Event studies: a review of issues and methodology, *Quarterly Journal of Business and Economics* 28, 36–67.

# Exhibit 61

# *p*-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results

Uri Simonsohn<sup>1</sup>, Leif D. Nelson<sup>2</sup>, and Joseph P. Simmons<sup>1</sup>

<sup>1</sup>University of Pennsylvania and <sup>2</sup>University of California, Berkeley

Perspectives on Psychological Science  
2014, Vol. 9(6) 666–681

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/1745691614553988

pps.sagepub.com



## Abstract

Journals tend to publish only statistically significant evidence, creating a scientific record that markedly overstates the size of effects. We provide a new tool that corrects for this bias without requiring access to nonsignificant results. It capitalizes on the fact that the distribution of significant *p* values, *p*-curve, is a function of the true underlying effect. Researchers armed only with sample sizes and test results of the published findings can correct for publication bias. We validate the technique with simulations and by reanalyzing data from the Many-Labs Replication project. We demonstrate that *p*-curve can arrive at conclusions opposite that of existing tools by reanalyzing the meta-analysis of the “choice overload” literature.

## Keywords

publication bias, *p*-hacking, *p*-curve

Scientific inquiry is concerned not only with establishing whether an empirical relationship holds, but also with estimating the size of that relationship. For example, policy makers not only want to know whether a particular policy will produce the desired effect, but also whether the size of that effect is large enough to justify putting the policy into action. To estimate effect sizes of particular relationships, scientists often conduct meta-analyses, combining the results of many similar studies into a single effect size estimate. Unfortunately, because of biases in the publication process, producing an accurate effect size estimate is often extremely difficult.

Scientific journals usually do not publish results unless they are statistically significant (henceforth assumed to correspond to  $p \leq .05$ ), a fact we will refer to as *publication bias* (see e.g., Fanelli, 2012; Rosenthal, 1979; Sterling, 1959). Because overestimated effect sizes are more likely to be significant than are underestimated ones, the published record systematically overestimates effect sizes (Hedges, 1984; Ioannidis, 2008; Lane & Dunlap, 1978).

To illustrate, imagine a researcher investigating whether people in a happy mood are willing to pay more for experiences than are people in a sad mood. She randomly assigns 40 people to watch either a happy video or a sad video and then measures their willingness to pay

for a ticket to see their favorite band in concert. With 20 people per condition, the two condition means would have to differ by at least .64 standard deviations (i.e.,  $\hat{d} \geq .64$ ) for them to be significantly different.<sup>1</sup> Thus, no matter what the true effect size is, with 20 observations per condition, the average significant effect size must be at least .64 standard deviations. In fact, even if an effect does not exist at all ( $d = 0$ ), the effect size estimated from just the significant studies will be large, with the means differing by  $\hat{d} = .77$  standard deviations (see Fig. 2A).

Scientists wanting to estimate the true size of an effect need to correct the inflated effect size estimates that publication bias produces. In this article, we introduce a new and better method for doing so. This method derives effect size estimates from *p*-curve, the distribution of significant *p* values across a set of studies (Simonsohn, Nelson, & Simmons, 2014). We show that this simple technique, which requires only that we obtain significant *p* values from published studies, allows scientists to much more accurately estimate true effect sizes in the presence

## Corresponding Author:

Uri Simonsohn, University of Pennsylvania - The Wharton School, 500 Huntsmann Hall, 3730 Walnut Street, Philadelphia, PA 19104  
E-mail: uws@wharton.upenn.edu



of publication bias. When the publication process suppresses nonsignificant findings,  $p$ -curve's effect size estimates dramatically outperform those generated by the most commonly used technique for publication-bias correction.

### **$p$ -Curve and Effect Size**

$p$ -curve is the distribution of statistically significant  $p$  values ( $p < .05$ ) across a set of studies.<sup>2</sup> For example, if four studies report critical  $p$  values of .043, .039, .021, and .057,  $p$ -curve for this set of studies would include all of those that are below .05: .043, .039, .021, but not .057. In a previous article, we showed how  $p$ -curve's shape diagnoses whether a set of studies contains evidential value or not (Simonsohn et al., 2014). In this article, we show how one can use  $p$ -curve's shape to estimate the average true effect size across the set of studies included in  $p$ -curve. Note that in the face of publication bias, the average true effect size will differ from the average observed effect size, and that  $p$ -curve will estimate the former. Here is an intuitive way to think of  $p$ -curve's estimate: It is the average effect size one expects to get if one were to rerun all studies included in  $p$ -curve.

A  $p$  value reflects the likelihood of observing at least as extreme an estimate if there is truly no effect ( $d = 0$ ). Thus, by definition, if an effect is not real, then 5% of  $p$  values will be below .05, 4% will be below .04, 3% will be below .03, 2% will be below .02, and 1% will be below .01. Thus, under conditions of no effect ( $d = 0$ ), there will be as many  $p$  values between .04 and .05 as between .00 and .01, and  $p$ -curve's expected shape is *uniform*.

If an effect exists, then  $p$ -curve's shape changes. Its expected distribution will be right-skewed: We expect to observe more low significant  $p$  values ( $p < .01$ ) than high significant  $p$  values ( $.04 < p < .05$ ; Cumming, 2008; Hung, O'Neill, Bauer, & Kohne, 1997; Simonsohn et al., 2014; Wallis, 1942). For any given sample size, the bigger the effect, the more right-skewed the expected  $p$ -curve becomes. Figure 1 shows some examples.

Conveniently, for a particular statistical test,  $p$ -curve's expected shape is solely a function of sample size and effect size. Because of this, knowing  $p$ -curve's shape and the sample size for a set of studies allows one to compute the effect size. Holding sample size constant, a greater proportion of small significant  $p$  values (i.e., a more extreme right skew) implies a larger effect size.

To get an intuition for how to estimate effect sizes using  $p$ -curve, consider a difference-of-means test with  $n = 20$  per cell. As shown in Figure 1, if the true effect size is  $d = .42$ , then 38% of significant  $p$  values are expected to be below .01. If the true effect size is  $d = .91$ , then 71% of significant  $p$  values are expected to be below .01. Thus, for a set of studies with  $n = 20$  per sample, if 38% of

significant  $p$  values are below .01, then our best guess of these studies' average effect size would be  $\hat{d} = .42$ . If 71% of significant  $p$  values are below .01, then our best guess of these studies' average effect size would be  $\hat{d} = .91$ .

More generally, for an observed set of significant results, one can identify the expected  $p$ -curve that most closely resembles the observed  $p$ -curve, and then identify the effect size estimate corresponding to that  $p$ -curve. Because the shape of  $p$ -curve is a function exclusively of sample size and effect size, and sample size is observed, we simply find the effect size  $\hat{d}$  that obtains the best overall fit. In the Appendix, we provide a detailed account (and R code) of how this is done.<sup>3</sup>

### **Selectively Reporting Significant Studies**

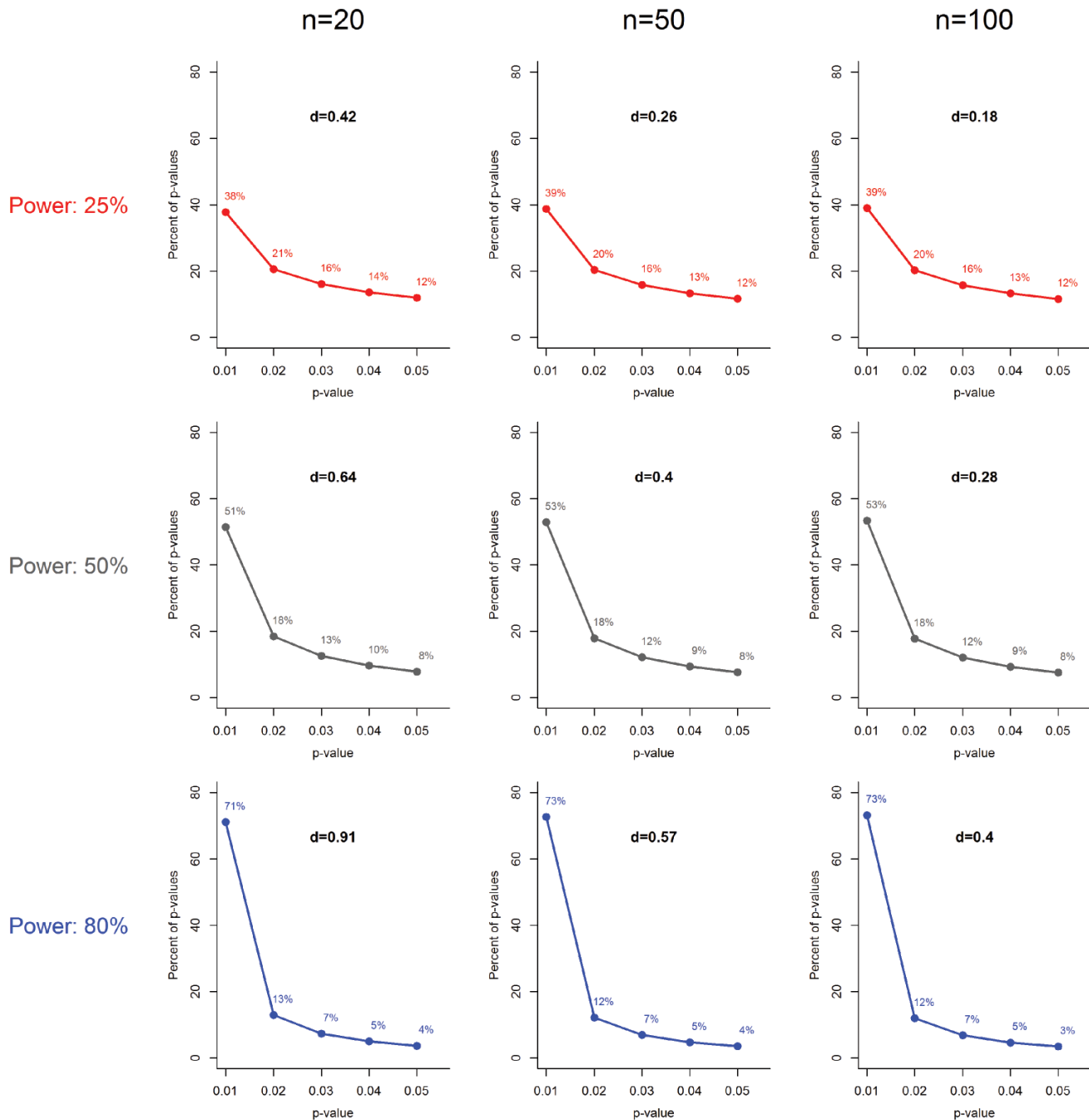
As described above, scientists interested in estimating effect sizes must correct for the fact that nonsignificant findings are much less likely to be published than are significant findings. This fact has long been recognized and a variety of corrective techniques have been proposed (for a review, see Rothstein, Sutton, & Borenstein, 2005).

The most common technique is known as *Trim and Fill* (Duval & Tweedie, 2000a, 2000b).<sup>4</sup> Although Trim and Fill is in common use, it rests on the unlikely assumption that the selective reporting of studies is driven by effect size rather than statistical significance.<sup>5</sup> That is, it assumes that the publication process suppresses the publication of small effects (regardless of significance) rather than nonsignificant results (regardless of effect size).<sup>6</sup>

However, publication bias in psychology (Rosenthal, 1979; Sterling, Rosenbaum, & Weinkam, 1995) and other fields (Ashenfelter, Harmon, & Oosterbeek, 1999; Gerber & Malhotra, 2008) primarily involves the suppression of nonsignificant results. As shown below, when the publication process suppresses nonsignificant findings, Trim and Fill is woefully inadequate as a corrective technique.  $p$ -curve performs much better.<sup>7</sup>

We conducted simulations to examine how well  $p$ -curve corrects for the selective reporting of statistically significant studies, and contrasted its performance with that of Trim and Fill. Specifically, we simulated studies testing a directional prediction with a two-sided difference-of-means test, pooled all the statistically significant studies with the predicted effect into a meta-analysis that included about 5,000 studies, and, to capture the selective reporting of significant studies, we estimated the true effect size based on the statistically significant studies alone.

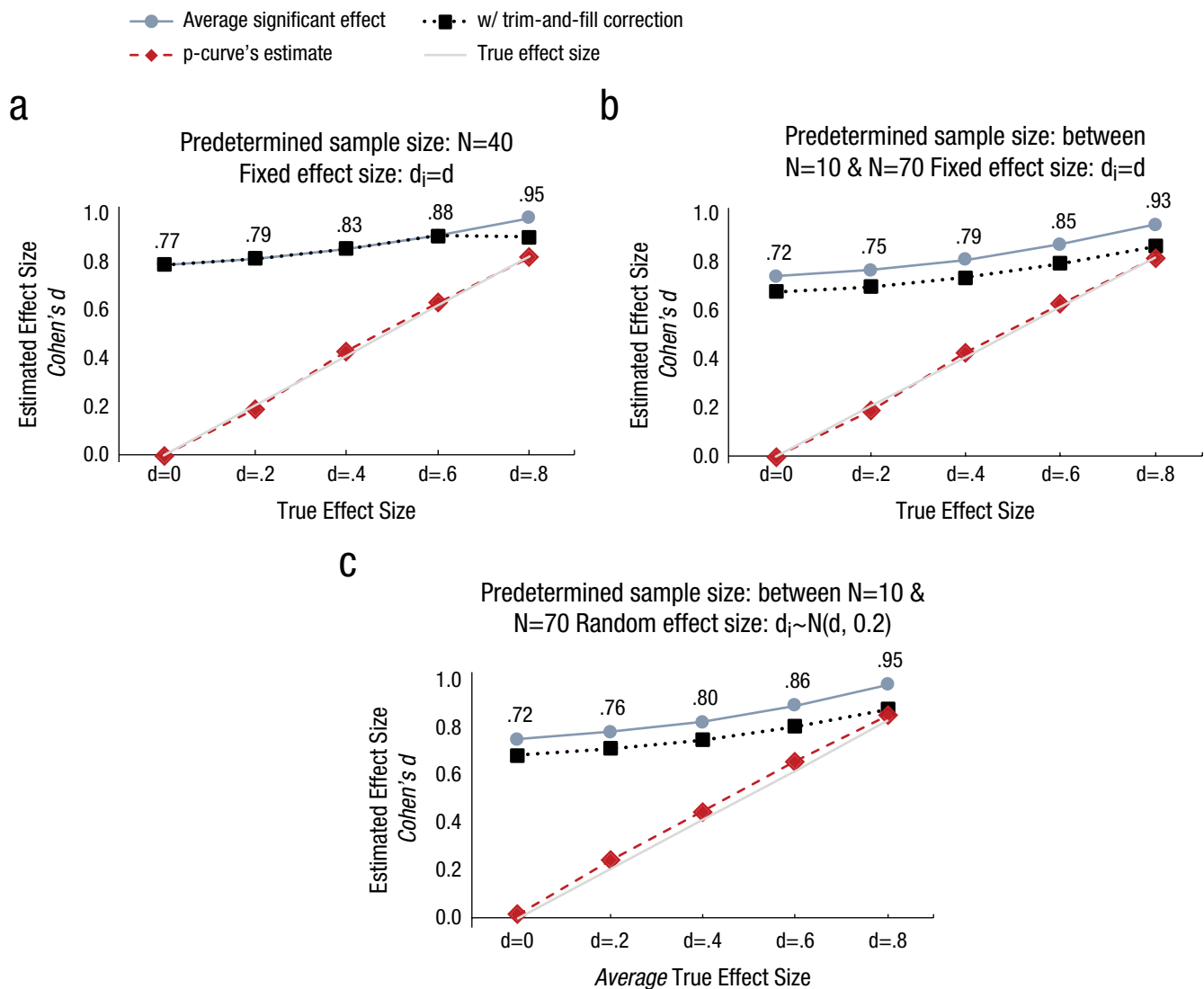
We estimated the effect size using three different approaches. First, we conducted a traditional fixed-effect



**Fig. 1.** *p*-curve's shape as a function of sample size and effect size. Expected *p*-curves for two-sample difference-of-means *t* tests, with *n* subjects per cell, where population means differ by *d* standard deviations from each other. Note that for a given level of power, *p*-curve is almost the same for every underlying sample-size and effect-size combination.<sup>20</sup> Plotted results are obtained from noncentral *t* distributions (see Appendix for details).

meta-analysis, computing the weighted average of observed effect sizes across the significant studies, without any correction for selectively reporting studies. Second, we corrected this estimate using the Trim and Fill procedure (Duval & Tweedie, 2000a, 2000b). Third, we estimated effect size using *p*-curve.

As shown in Figure 2, we conducted these simulations under a number of conditions. Panel A reports results assuming that all studies within a meta-analysis have the same sample size ( $n = 20$  per cell) and effect size (shown on the *x*-axis). Panel B reports results allowing for studies within a meta-analysis to vary in sample size between



**Fig. 2.** Impact of selectively reporting significant studies. Each marker reports an effect size estimate (Cohen's  $d$ ) based on a meta-analysis performed on about 5,000 statistically significant simulated studies. In Panel A, all studies have 20 observations per cell and assume a fixed effect size. In Panel B, the same number of statistically significant studies with each sample size between  $n = 5$  and  $n = 35$  per cell are included in the meta-analysis. Panel C is the same as Panel B, except effect sizes are drawn from a normal distribution with mean  $d$ , standard deviation  $\sigma = .2$ . Trim and Fill and  $p$ -curve estimates are based exclusively on  $p < .05$  results.

$n = 5$  and  $n = 35$  per cell. In Panel B's simulations, all studies within a meta-analysis had the same true effect size, and each meta-analysis included the same number of studies with each sample size.

The simulations in Panel C varied both sample size and true effect size across studies included in the same meta-analysis. For each simulated study, we first randomly drew a true effect size from a normal distribution with  $\sigma = .2$  and the mean indicated on the graph's  $x$ -axis. We then drew observations from populations whose true means differed by that random effect while varying per-cell sample size to be between  $n = 5$  and  $n = 35$ , and we pooled the set of statistically significant results. All

studies within a meta-analysis had the same average true effect size, and each meta-analysis included the same number of statistical significant studies with each sample size.

Figure 2 displays the results, revealing several important facts. First, it shows the dramatic inflation of effect size generated by publication bias (Hedges, 1984; Ioannidis, 2008; Lane & Dunlap, 1978). The darker solid lines in the figure show that the average effect sizes of the subset of studies that are statistically significant are dramatically higher than the true effect sizes. Moreover, the estimates hardly vary as a function of true effect size. When samples are small, the subset of statistically

significant effect sizes contains almost no information about the true effect size.

Second, applying the Trim and Fill correction to these estimates does not make them meaningfully better. For example, when there was truly no effect ( $d = 0$ ), Trim and Fill estimated the effect to be large, at least  $d = .65$ . When nonsignificant studies are not observed, the most popular corrective technique is not very corrective.

Third, in contrast to the other methods,  $p$ -curve fully corrects for the impact of selectively reporting studies. For example, in Panel A, we see that when the true effect size is  $d = 0$ ,  $p$ -curve correctly estimates the effect to be zero, despite  $p$ -curve being based exclusively on observed estimates of  $d > .64$ , with an average observed effect size of  $\hat{d} = .77$ . Panels B and C show that the accuracy of  $p$ -curve does not rely on homogeneity of sample size nor effect size. In all cases,  $p$ -curve is accurate and the other methods are not.<sup>8</sup>

The results from Figure 2 assess the performance of Trim and Fill when performed only on the subset of statistically significant findings, showing that it provides a minimal improvement over the naive average effect size. In Supplement 3, we show that adding nonsignificant findings to the set analyzed via Trim and Fill, even doubling the total number of studies, does not noticeably improve the corrective abilities of Trim and Fill.

## **$p$ -Hacking**

Researchers not only selectively report studies, they also selectively report analyses within a study (Cole, 1957; Simmons, Nelson, & Simonsohn, 2011). For example, a researcher may run a regression with and without outliers, with and without a covariate, with one and then another dependent variable, and then only report the significant analyses in the paper. We refer to this behavior as  $p$ -hacking (Simmons, Nelson, & Simonsohn, 2012; Simonsohn et al., 2014).

$p$ -hacking enables researchers to find statistically significant results even when their samples are much too small to reliably detect the effect they are studying or even when they are studying an effect that is nonexistent. For this reason, existing methods for estimating effect sizes will be inflated, often dramatically, in the presence of  $p$ -hacking.

$p$ -hacking biases  $p$ -curve's effect size estimates as well, but it does so in the opposite direction, leading one to underestimate effect sizes. To understand why, consider the effects that  $p$ -hacking has on  $p$ -curve's shape.

Because  $p$ -hacking leads researchers to quit conducting analyses upon obtaining a statistically significant finding,  $p$ -hacking is disproportionately likely to introduce "large" significant  $p$  values into the observed distribution (i.e.,  $p$  values just below .05). As a result,  $p$ -hacking reduces the right skew of  $p$ -curve (Simonsohn et al.,

2014). Because smaller effect sizes are associated with less right-skewed  $p$ -curves,  $p$ -hacking causes  $p$ -curve to underestimate effect sizes.

To explore the effects of  $p$ -hacking on effect size estimates, we simulated three common forms of  $p$ -hacking (John, Loewenstein, & Prelec, 2012; Simmons et al., 2011): achieving statistical significance by (a) data peeking (collecting more observations if an initial sample of observations does not obtain  $p < .05$ ), (b) selectively reporting which of three dependent variables to report, and (c) selectively excluding outliers.

All of the simulations explored mean differences between two conditions starting with 20 observations in each. We varied the true effect size across simulations. For each simulation, we estimated effect size in three ways: (a) by computing the average of all effects (including all significant and nonsignificant findings), (b) by applying a Trim and Fill correction to only the statistically significant effects, and (c) by using  $p$ -curve. We report further details of these simulations in the next section; readers may choose to skip ahead to the Results section.<sup>9</sup>

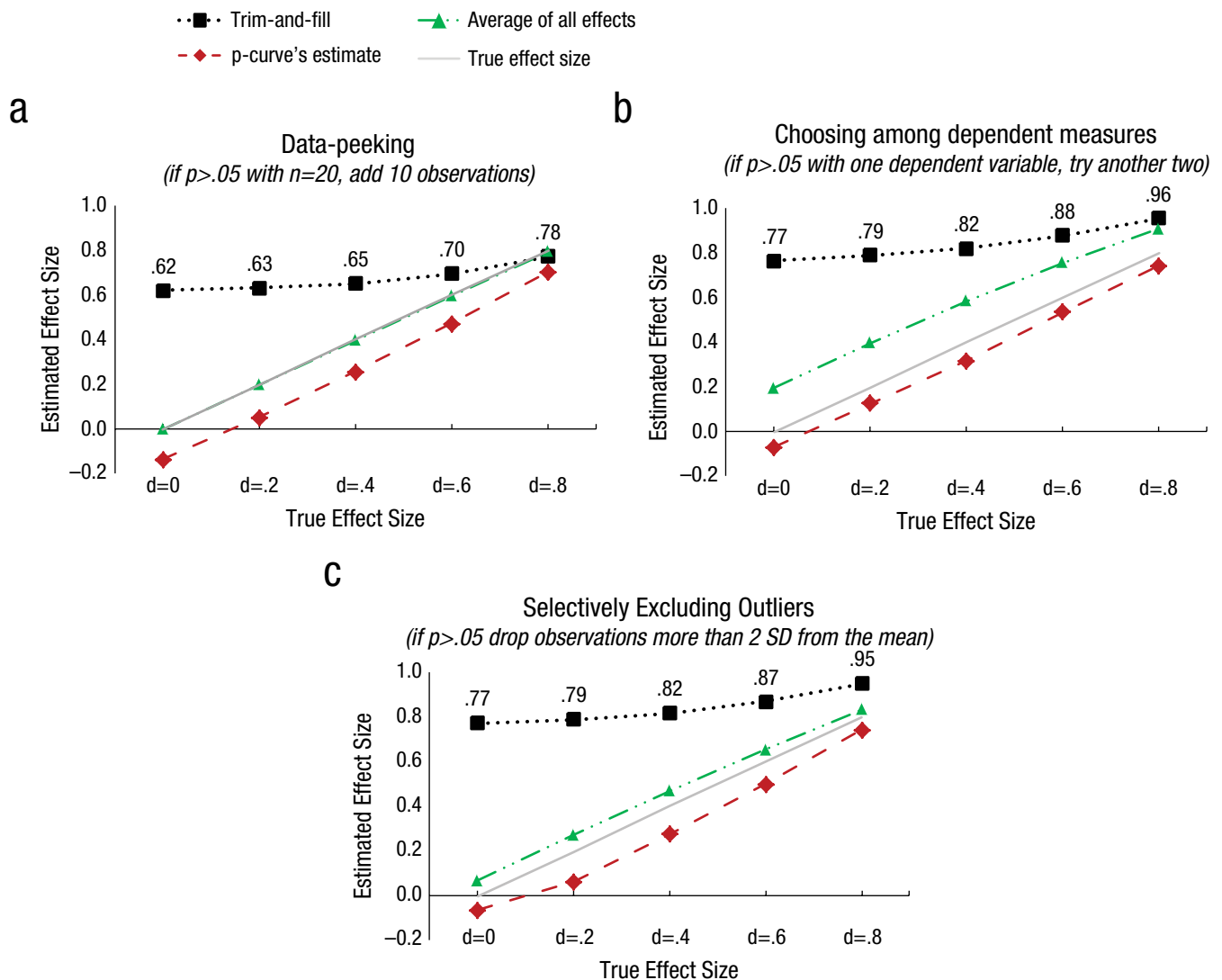
## **Details**

Figure 3A shows the results of simulations of data peeking. For each study, we first conducted a  $t$  test with  $n = 20$  observations per cell. If the result was significant, we "published" it; if it was not, we added 10 observations to each sample, thus increasing the per-condition sample size from 20 to 30, and conducted a new  $t$  test. If this second result was significant, we "published" it; if not, it remained nonsignificant and hence "unpublished."

Figure 3B shows the results of simulations of selectively reporting among three dependent variables correlated with each other at  $r = .5$ . We conducted a  $t$  test on each of these dependent variables. Within a study, the first comparison to obtain significance was "published"; if all three  $t$  tests were nonsignificant, the study was "unpublished," and the analysis yielding the lowest  $p$  value was the one used to compute the average effect across all of the studies.

Figure 3C shows the results of simulations of selectively dropping outliers further than two standard deviations away from the sample mean. For each study, we conducted four  $t$  tests: one comparing both full samples, one dropping outliers from only the first sample, one dropping outliers from only the second sample, and one dropping outliers from both samples. Within a study, the first  $t$  test to obtain significance was "published." If all four  $t$  tests were nonsignificant, the study was "unpublished," and the analysis yielding the lowest  $p$  value was the one used to compute the average effect across all of the studies.

For each of these simulated forms of  $p$ -hacking we pooled all the significant results and estimated the underlying effect size using the three methods described above:



**Fig. 3.** Impact of  $p$ -hacking. Each marker reports an effect size estimate (Cohen's  $d$ ) based on a meta-analysis performed on large numbers of studies. The triangle dash-dot line plots estimates based on all simulated studies (what would be estimated by a meta-analyst that obtained all studies ever conducted), and the square dotted line and diamond dashed line plot estimates based on the statistically significant subset. Each panel simulates a different form of  $p$ -hacking. A: Adding 10 observations per-cell if  $p > .05$  is not achieved with  $n = 20$ . B: Analyzing three different dependent variables, reporting either the first to yield  $p < .05$ , or, if none are significant, the lowest  $p$  value obtained. C: Excluding observations further than two standard deviations from the condition's mean—first only from one condition, then the other, then both—if  $p > .05$  has not yet been achieved. The square dotted lines are based on the last analysis conducted and hence may include bias from  $p$ -hacking; if they were based on the first, they would be free of  $p$ -hacking and hence always identical to the gray solid lines. Trim and Fill and  $p$ -curve estimates are based exclusively on  $p < .05$  results.

the average of all studies regardless of significance, the estimate derived from applying the Trim and Fill correction to the significant studies, and  $p$ -curve's effect size estimates.

## Results

Figure 3 shows the results. First, we again see that Trim and Fill does not adequately estimate effect sizes when the publication process suppresses nonsignificant results. Second, we see that these forms of  $p$ -hacking cause the

$p$ -curve to underestimate effect sizes. Interestingly,  $p$ -hacking biases not only the published record, but the totality of evidence, which includes all studies whether or not they are likely to be published. Thus, even if a meta-analyst were to gain access to every single study, eliminating publication bias via “brute force,”  $p$ -hacking will cause effect sizes to be overestimated.

Thus,  $p$ -hacking will bias effect size estimates regardless of how effect sizes are estimated. Using traditional methods,  $p$ -hacking will bias effect sizes upwards; when using  $p$ -curve,  $p$ -hacking will bias effect sizes

downwards. Because the relative magnitude of these biases is situation specific, it is not possible to make general statements as to whether analyzing all studies ever conducted would outperform  $p$ -curve's estimate based only on the statistically significant subset.

Note that the results from Figure 3 assume that unpublished results remained  $p$  hacked (e.g., that observations that were dropped by a researcher but did not succeed in lowering the  $p$  value to  $p < .05$ , would also be excluded from the dataset given to the meta-analyst). If unpublished results were free of  $p$ -hacking—for example, if any dropped observations were reintroduced before the data were meta-analyzed—then the bias present in the meta-analysis of all conducted results would be reduced. It would still not be eliminated, because the published studies included in the meta-analysis are still biased upwards by  $p$ -hacking.

## Precision

We have so far reported results from simulations involving large numbers of studies. We now turn to the issue of how much precision we may expect from  $p$ -curve's estimates when relying upon smaller sets of studies. Figure 4 reports results from simulations that varied true effect sizes, the number of studies included in  $p$ -curve, and the sample sizes of those studies (with per-cell sample size of either 20 or 50). The markers indicate the median effect size estimate across simulations, and the vertical bars indicate one standard error above and below that median (i.e., the standard deviation of that estimate across simulations).

As one may expect,  $p$ -curve is more precise when it is based on studies with more observations and when it is based on more studies. Less obvious, perhaps, is the fact that larger true effects also lead to more precision. This occurs because  $p$ -curve's expected shape quickly becomes very right-skewed as effect size increases, reducing the variance in skew of observed  $p$ -curves.

## Demonstrations

In this section, we provide two demonstrations. The first relies on data from the “Many-Labs replication project” (Klein et al., 2014), in which 36 different labs around the world collaborated to run the exact same set of studies and reported all results regardless of statistical significance. This provides a unique opportunity to assess the performance of  $p$ -curve in a realistic environment—using real studies, real dependent variables, and real participants—where we nevertheless observe all studies conducted, regardless of outcome.

The second demonstration revisits the meta-analysis of the popular psychology literature on choice overload

(Scheibehenne, Greifeneder, & Todd, 2010). This example demonstrates a situation in which  $p$ -curve and traditional meta-analytical tools arrive at different answers, suggesting different paths for what future empirical work on the topic ought to seek to accomplish.

### **Demonstration 1. Many Labs Replication Project**

Klein et al. (2014) conducted replications examining 13 different “effects” across 36 labs (data available from <https://osf.io/wx7ck/>). We can use these data to assess how well  $p$ -curve corrects for publication bias in a realistic setting by comparing the effect size estimate we obtain from  $p$  curving only the subset of significant results to that obtained by averaging the results from all labs.

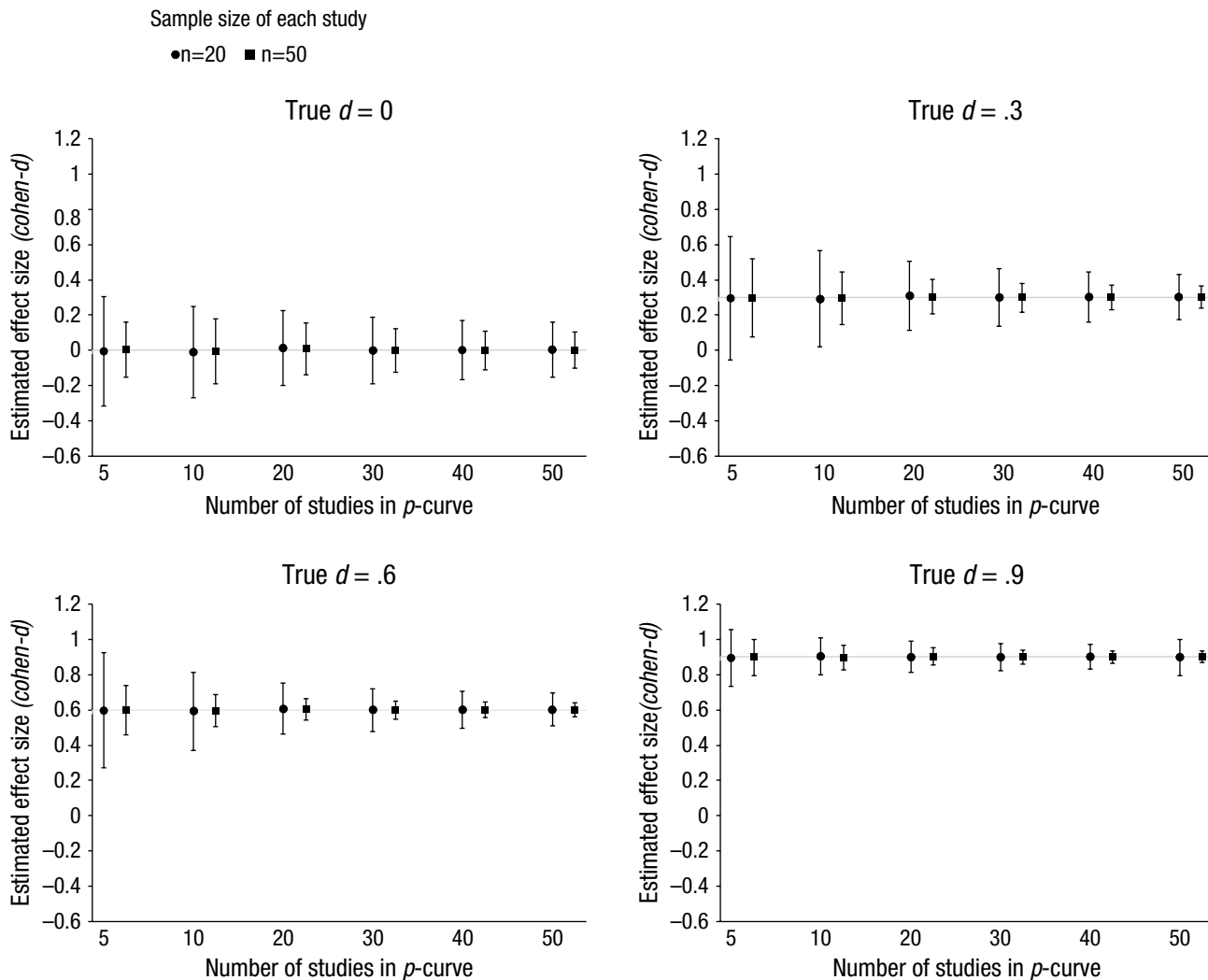
For this assessment of performance to make sense, and for the aggregate average to be a valid proxy for truth, we need to believe that the studies that worked and did not work were examining the same average effect—that they differed only because of sampling error. Otherwise, if  $p$ -curve recovers one estimate, and the aggregate average another, we don't know if  $p$ -curve performed poorly, or if it is correctly indicating that the significant and nonsignificant studies were investigating a different underlying effect. We thus focus on effects that proved homogeneous across the different labs.<sup>10</sup>

The two most homogenous effects were the sunk costs fallacy (as studied by Oppenheimer, Meyvis, & Davidenko, 2009) and the Asian disease problem (Tversky & Kahneman, 1981).<sup>11</sup> Conveniently, these two effects were associated with very different replication rates. Only 50% of labs obtained a significant result for the sunk cost fallacy, the lowest in the set of effects deemed “replicable,” whereas 86% of the studies investigating the Asian disease problem were significant.<sup>12</sup>

Figure 5A shows the resulting  $p$ -curves. Both are right skewed, but Asian disease's  $p$ -curve was more so. Whereas 83% of the Asian Disease Problem's significant  $p$  values were below .01, only 31% of the Sunk Cost Fallacy's significant  $p$  values were below .01. Figure 5B reports the resulting effect size estimates, comparing  $p$ -curve's estimates to a naive estimate, computed by averaging the effect size observed across the significant studies, and an earnest estimate, computed by averaging the effect size across all studies, regardless of significance. Because these results were not  $p$  hacked, we can safely assume that the earnest estimate represents the best estimate of the true effect size.<sup>13</sup>

The bias of the naive estimate is small for the Asian disease problem, as a large proportion of those studies were significant. It estimates a true effect size of .66, whereas the average across all studies was .60. Reassuringly,  $p$ -curve's estimate agrees with the earnest





**Fig. 4.** Precision of  $p$ -curve. Each marker reports the median effect size obtained across 1,000 simulations of meta-analyses including between 5 and 50 studies, each with the same sample size ( $n = 20/50$  per cell) and underlying true effect. Vertical bars show one standard error above and one standard error below the estimate. Standard errors are computed as the standard deviation of the parameter estimate across simulations.

estimate, and thus corrects little when little needs to be corrected. The bias of the naive estimate for the sunk cost fallacy is much larger, estimating a true effect size of .46 when the average across all studies was .31. Reassuringly,  $p$ -curve's estimate again agrees with the earnest estimate and thus corrects more when more needs to be corrected.<sup>14</sup>

### Demonstration 2. Choice overload

The choice overload literature in psychology examines whether an increase in the number of options available leads to negative consequences, such as a decrease in motivation to choose or satisfaction with the option that is ultimately chosen. Scheibehenne et al. (2010) conducted a meta-analysis combining published and unpublished studies, obtaining an overall mean effect size of

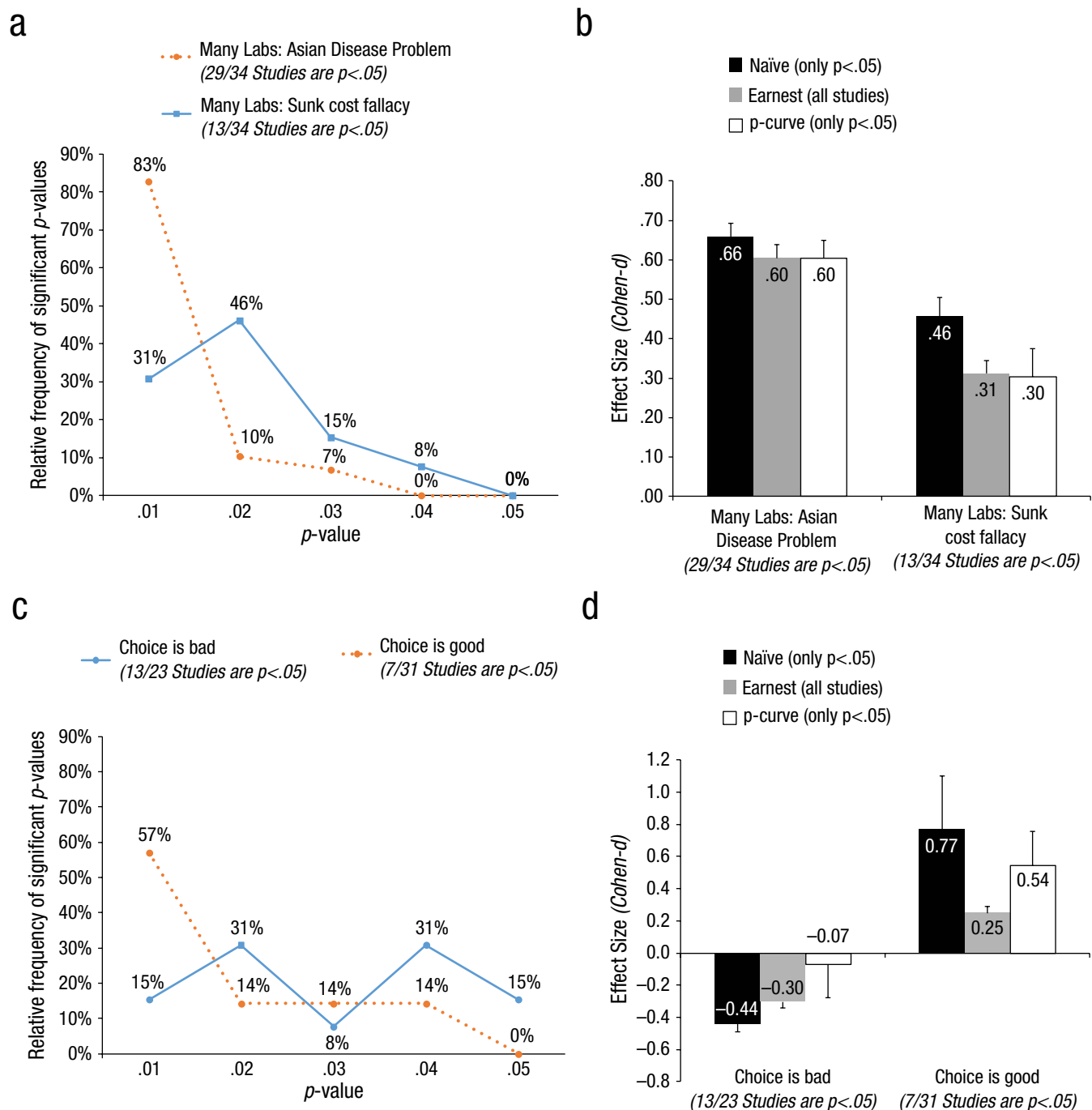
“virtually zero” (p. 409). This result implies that changes in choice-set size are inconsequential.

As Chernev, Bockenholt, and Goodman (2010) commented:

“... studies often include two conditions: one designed to show that the effect of the construct (e.g., choice overload) is present, *and another one designed to document the directionally opposite* (e.g., more-is-better) effect [...] combining their effect sizes to test their average effect leads to a potentially biased interpretation of the underlying effects.” (p. 427, emphasis added)

In line with their (correct, in our view) observation, we split the studies analyzed by Scheibehenne et al. into two groups: one showing “choice is good” (positive





**Fig. 5.** Demonstrations of  $p$ -curve. Panel A depicts  $p$ -curves for studies reported in the Many Labs replication project (Klein et al., 2014), and Panel C depicts  $p$ -curves for studies included in the meta-analysis of the impact of choice-set size on consumer outcomes (Scheibehenne, Greifeneder, & Todd, 2010). The latter are split into studies showing a positive effect and a negative effect, as suggested by a commentary on that meta-analysis (Chernev, Bockenholt, & Goodman, 2010). Panels B and D depict the corresponding effect size estimates obtained via traditional meta-analytical tools applied only to statistically significant studies (naïve), to all studies available (earnest), and from applying  $p$ -curve to the significant subset. Vertical bars correspond to one standard error.

coefficient of choice set size), and the other showing “choice is bad” (negative coefficient). We estimated effect sizes separately for both sets of studies, effectively asking two conceptually and statistically independent questions:

- (1) In studies showing that more choice is good, how good is choice?
- (2) In studies showing that more choice is bad, how bad is choice?

The results are reported in Figure 5D. When we limit our estimation to statistically significant studies and average across them, both effect size estimates are sizeable; both get closer to zero when we add the nonsignificant studies. Interestingly, when we apply *p*-curves to the significant findings, the choice-is-bad effect gets smaller, becoming effectively zero, whereas the choice-is-good effect gets larger. The right skewed *p*-curve for choice-is-good and flat *p*-curve for choice-is-bad, depicted in Figure 5C, reveal why *p*-curve estimates move in opposite directions for the two sets of studies.

Some of the error bars in Figure 5D are large. Particularly relevant is the standard error for the  $\hat{d} = -.07$  estimate for choice-is-bad. *P*-curve's estimate is not conclusively saying the effect is 0; it is saying the best guess is close to zero, but that the data are consistent with practically and theoretically relevant effect sizes (of both signs) also.

Our interpretation of Figure 5D is the following. If we conduct traditional meta-analysis on the available evidence, we empirically verify the concern expressed by Chernev et al. (2010): under predictable circumstances choice is good, and under predictable circumstances choice is bad. This would suggest that future work in the choice overload literature could safely and constructively consist of conceptual replications that seek to learn more about moderators and mediators of the choice-is-bad effect.

If we conduct meta-analysis based on *p*-curve, the conclusion is different. Yes, the evidence does support the (uninteresting) notion that under predictable circumstances choice is good, but it neither confirms nor denies the (much more interesting) notion that under predictable circumstances choice is bad. The conclusion based on *p*-curve suggests that future work should focus on examining if the basic phenomenon of choice-is-bad can indeed be reliably obtained. Properly powered direct replications of the original demonstrations are something top journals may be less inclined to publish based on the traditional meta-analysis result ("We already know this!") but more inclined to publish based on our *p*-curve results ("We don't really know the answer to this yet"). Thus, given the same data, *p*-curve and traditional meta-analysis can suggest very different paths forward.

## Limitations

### **Limitation 1. Simple effects from attenuated interactions cannot be analyzed**

The validity of *p*-curve rests on the assumption that a lower *p* value does not increase the likelihood of publication once it crosses the significance threshold. For example, a significance criterion of .05 assumes that a *p* value

of .008 would have been publishable if it had instead been .038.

Though this assumption usually holds, studies investigating attenuated interactions often violate it. An attenuated interaction hypothesis is one that predicts that an effect will be smaller under one condition than under a different condition. For example, the hypothesis that the effect of gender on height will be smaller for children than for adults is an attenuated interaction hypothesis. In this example, the *unattenuated simple effect* is the (larger) effect of gender on height for adults, whereas the *attenuated simple effect* is the effect of gender on height for children.

Researchers interested in publishing attenuated interactions need the interaction terms to be significant (Gelman & Stern, 2006). But for the interaction term to be significant, the (larger) unattenuated simple effect needs to have an even lower *p* value. For example, if the simple effect of gender on adults' heights was associated with  $p = .038$ , it is unlikely that the interaction would be significant. This means that the de facto significance criterion for unattenuated simple effects in this design is smaller than for the other *p* values. As a result, the inclusion of *p* values for unattenuated simple effects will result in a *p*-curve that overestimates effect sizes. The inclusion of only the interaction term *p* value would leave *p*-curve unbiased. As a result, for studies hypothesizing attenuated interactions, we recommend never including results from simple effects in *p*-curve.

It is worth noting that this problem does not apply to studies predicting reversing interactions, involving an effect being observed under one condition but then the opposite effect being observed under a different condition. In this case, both simple effects may be included in *p*-curve.

### **Limitation 2. *p*-curve ignores nonsignificant results**

Another limitation is that *p*-curve ignores information from nonsignificant studies ( $p > .05$ ). Because this limits the sample size of studies under consideration, *p*-curve is more likely to provide a noisy effect size estimate. This is a necessary limitation of *p*-curve: Because we do not know what publication pressures occur above .05, we do not know what the distribution of *p* values should be above .05. Note that although excluding nonsignificant results makes *p*-curve noisier (less efficient), it does not make *p*-curve biased.

### **Limitation 3. Downward bias with *p*-hacking**

As we discussed in some detail when presenting the results from Figure 3, *p*-hacking can bias effect size

estimates from  $p$ -curve downwards. In our simulations, however, this bias is mild enough to be ignorable.

#### **Limitation 4. Moderation within meta-analysis**

In its current form  $p$ -curve estimates a single average for all studies included in it. To examine if a variable moderates the effect size of interest across studies, then separate  $p$ -curves would need to be estimated. For example, to examine if anchoring studies run in the lab have different effect sizes than anchoring studies run outside the lab, one would perform one  $p$ -curve for lab studies and one  $p$ -curve for nonlab studies, rather than a single  $p$ -curve that includes a moderator variable. It seems likely that  $p$ -curve can be modified to incorporate moderation within a single analysis, but we have not explored that possibility.

#### **Another Use of $p$ -curve: Average Power**

$p$ -curve's shapes is closely tied to statistical power, the probability that a study obtains a significant result. For a given statistical test, both power and  $p$ -curve depend exclusively on the size of the sample and the size of the effect. This means that  $p$ -curve can be used to estimate the average underlying statistical power of a set of studies. As with effect sizes,  $p$ -curve's estimate of power will correct for the inflated estimates that arise from the privileged publication of significant results.

Estimating the publication-bias corrected estimate of the average power of a set of studies can be useful for at least two purposes. First, many scientists are intrinsically interested in assessing the statistical power of published research (see e.g., Button et al., 2013; Cohen, 1962; Rossi, 1990; Sedlmeier & Gigerenzer, 1989). But to carry out their calculations they have either (a) relied on arbitrary effect size assumptions (e.g., small, medium, and large) and asked how much power do the observed sample sizes have to detect effects of that size, or (b) computed the average observed effect size (which is inflated by publication bias and causes problems if effect size is heterogeneous) and computed the post-hoc power for that effect. With  $p$ -curve, those arbitrary assumptions are no longer needed, and we can estimate the actual underlying power correcting for publication bias.

Second, published results often provide insufficient details to compute effect size. For example, effect size in mixed between-/within-subject designs depends on within-participant correlations across observations—a metric that is often not reported. Nevertheless, the true underlying power of the reported test statistic (e.g.,  $F$  test in an analysis of variance) and hence the resulting  $p$ -curve is of course influenced by that parameter whether it is

reported or not and, hence, so is the average power estimated via  $p$ -curve. When aggregating results reported in insufficient detail, one may estimate average underlying power and then convert the result into an intuitive metric of effect size taking into account the underlying sample size.

#### **User Guide**

Upon identifying the effect of interest (e.g., the impact of a high vs. low anchor value on monetary valuations) and the selection criterion for studies to include (e.g., all studies that cite Chapman and Johnson, 1999, and use a monetary dependent variable), the researcher must identify, for each study, the test statistic associated with testing the null that the effect of interest is zero. The set of all such tests is then submitted to a  $p$ -curve analysis, either using the R Code included in this article or the online web-app available at [www.p-curve.com](http://www.p-curve.com).

$p$ -curve assumes these tests are statistically independent from one another. If multiple results from the same participants are reported in a paper, only one of them may be included in a given  $p$ -curve. If only one of the results is pertinent to the meta-analysis then the decision is easy: Only the pertinent test is selected.

For example, if an anchoring study presented participants with both a high and a low anchor and then asked them to separately value both a coffee mug and a pen, then a meta-analyst only concerned with the impact of anchors on mugs would only include the mug result.

If multiple results are pertinent to the meta-analyst's question, then a decision must be made. For example, if both the pen and mug were relevant to the question of interest to the meta-analysis, then  $p$ -curve may include either the test statistic associated with an analysis of the mug, with an analysis of the pen, or with an analysis of a composite of the mug and the pen (e.g., the averaged or summed ratings of the mug and pen). But  $p$ -curve can never include results from more than one of these analyses. Importantly,  $p$ -curve should also never include the average  $p$  value of multiple tests. For example, if the anchoring effect for coffee mugs was  $t(38) = 2.12$ ,  $p = .04$  and the anchoring effect for pens was  $t(38) = 2.43$ ,  $p = .02$ ,  $p$ -curve should not include the average  $p$  value,  $p = .03$ , nor the  $p$  value associated with the average test statistic,  $p = .028$ . It should include either  $p = .02$  or  $p = .04$ .<sup>15,16</sup>

When choices among multiple tests must be made, we recommend adhering to a prespecified selection rule (e.g., the test reported first) and then computing and reporting the result obtained under a different rule (e.g., the test reported last).

As implemented here, all test results entered into  $p$ -curve are assumed to be either examining effects of the

same sign (e.g., all significant effects in *p*-curve show that anchoring occurs, and none show that the reverse of anchoring occurs) or that the sign of the effect is not relevant (e.g., when computing average statistical power across heterogeneous findings). If neither of these conditions is met, two separate *p*-curves should be conducted, one for positive effects and one for negative ones.<sup>17</sup>

For scientific results to be interpretable, it is imperative that researchers disclose how they resolved ambiguities surrounding the collection and analysis of data (Simmons et al., 2011). For *p*-curve users in particular, this is easily achieved by supplementing their explicit identification of a study selection rule with a *p*-curve disclosure table (Simonsohn et al., 2014).

To examine the impact of a discrete moderator on effect size, the simplest way to proceed is to split up the studies into different subgroups (e.g., studies performed in the United States make up one subgroup, studies performed outside the United States make up another) and then *p*-curve is applied separately to each group, obtaining a separate effect size estimate for each subgroup. In its current implementation, *p*-curve does not allow for the analysis of continuous moderators (see the Limitations section).

## Overview of Supplementary Materials

The Supplementary materials include the following sections.

1. Robustness of *p*-curve to data that are not normally distributed
2. Robustness of *p*-curve to heterogeneity of effect size
3. Trim-and-Fill performance when some  $p > .05$  are observed
4. Alternative loss functions (to the one from the appendix)
5. R-Code for every result in this article (also available here: <http://www.p-curve.com/Supplement/Rcode>)

## Conclusions

The selective publication of significant studies and analyses leads the published record to overestimate the size of effects. We have shown that one can use the distribution of significant *p* values, *p*-curve, to easily and effectively estimate effect sizes that correct for the selective reporting of studies, vastly outperforming the most commonly used alternative, Trim and Fill. *p*-curve also outperforms existing methods when researchers selectively report analyses—when they *p* hack—but for many forms of *p*-hacking it underestimates true effect sizes.

However, the presence of *p*-hacking biases all known methods of effect size estimation, even when one averages across every study ever conducted. Overall, *p*-curve seems to be the best tool for estimating effect size when the publication process predicts statistically significant results.

## Technical Appendix: Approach and Algorithm for Estimating Effect Size Using *p*-curve

The goal is to determine the underlying effect size that leads to an expected *p*-curve that best fits the observed *p*-curve. This requires three steps: (a) linking underlying effect size with expected *p*-curves; (b) defining a *loss function*, a metric of how well a given expected *p*-curve fits the observed *p*-curve; and (c) finding the effect size that minimizes that loss function.<sup>18</sup>

### A1. Linking effect size with expected *p*-curve

We assume here familiarity with noncentral distributions. For most readers that's a terrible assumption, but it can be remedied by consulting Supplement 1 in Simonsohn et al. (2014).

For simplicity, we focus on independent two-sample difference of means *t* tests performed on samples of the same size (*n*). As an introduction, let's go over how we constructed Figure 1. The top left panel shows the expected *p*-curve when  $n = 20$  and  $d = .4164$  (shown as .42). To obtain that expected *p*-curve, we first identify the critical *t* values that lead to  $p = .01, .02, \dots, .05$  with a degree of freedom of 38.

Using R syntax, this involves:

```
x5 = qt(.975,df = 38)
x4 = qt(.98, df = 38)
x3 = qt(.985,df = 38)
x2 = qt(.99, df = 38)
x1 = qt(.995,df = 38)
```

For example,  $x_5 = 2.0244$ , which means that  $t(38) = 2.0244$  leads exactly to  $p = .05$  (for a two-sided test).

Next, we rely on the noncentral student distribution to see how likely one is to obtain *t* values more extreme than each of  $x_1, x_2, \dots, x_5$ . For the parameters above,  $n = 20$ ,  $\delta = .4164$ . This involves using the noncentrality parameter  $ncp = \sqrt{\frac{n}{2}} \delta = \sqrt{\frac{20}{2}} .4164$

Starting with  $x_5$  (again, the lowest *t* value that leads to a statistically significant result), we compute, in R-syntax again)

```
1-pt(x5,df=38,ncp=sqrt(20/2)*.4164) =.25
```

That is the probability of obtaining  $t \geq 2.0244$ , and hence  $p \leq .05$  when  $n = 20$  and  $\delta = .042$ —it is, hence, the statistical power of the test. There is a 25% chance of obtaining a statistically significant result with a study of those characteristics. The top-left panel of Figure 1 identifies what share of 25% of tests will be  $p < .01$ ,  $.01 < p < .02$ , etc. Let's determine what share of 25% of tests will be  $p < .01$ :

$$1 - \text{pt}(x1, df=38, ncp=\text{sqrt}(20/2) \cdot .4164)$$

We get: .094

This means that with  $n = 20$ , there is a 9.4% chance of obtaining  $p < .01$  when  $\delta = .42$ .

Among all attempted studies, 25% are  $p < .05$  and 9.4% are  $p < .01$ . Therefore, the share of significant results that are  $p < .01$  is 38% ( $9.4/25$ ; see Fig. 1). Proceeding analogously with  $x_2$ ,  $x_3$ , and  $x_4$ , we obtain the rest of the plotted numbers: the histogram version of  $p$ -curve.

For estimating effect size we treat  $p$ -curve as the continuous distribution that it is. We proceed analogously, but we do not limit ourselves to the five discrete points  $x_1$ – $x_5$ ; instead, we create a function that maps every possible statistically significant  $t$  value to the probability of obtaining a  $t$  value at least as large. This is effectively the  $p$  value of the  $p$  value, which we referred to as the *pp-value* (Simonsohn et al., 2014).

For  $t$  value  $t_i$ , the probability of observing at least as large a significant  $t$  value is:

$$\text{pp}(t_i) = \text{prob}(t > t_i \mid df, ncp, p < .05)$$

It is useful to get  $p < .05$  out of the conditional. Let's define,<sup>19</sup>

$$\text{power} = \text{prob}(p < .05 \mid df, ncp).$$

Then

$$\text{pp}(t_i) = (\text{prob}(t > t_i \mid df, ncp) - \text{power}) / \text{power}$$

$\text{pp}(t_i)$ , then, is the *cumulative distribution function* (*c.d.f.*) of  $p$ -curve, a function mapping sample and effect size

onto expected  $p$ -curve. Because we observe sample size, it effectively maps effect size onto expected  $p$ -curve.

#### A2. Defining a loss function

For every candidate effect size  $d_i$ , then, there is a  $\text{pp}(t \mid d_i)$  function that gives every possible  $t$  value a probability of observing at least as extreme a value. When  $d_i = \delta$  (when the candidate effect size equals the true effect size),  $\text{pp}$  values will be distributed uniformly for reasons entirely analogous to why  $p$  values are distributed uniformly under the null (which we explain in the main text).

In light of this, we define how well a given candidate effect size  $d_i$  fits the data—how well the expected  $p$ -curve fits the observed  $p$ -curve—by assessing how close to a uniform distribution the set of observed  $\text{pp}$  values are. Many techniques exist to compare empirical with expected distributions—we rely on the robust and simple Kolmogorov-Smirnov (KS) statistic, which computes the maximum observed gap between the two *c.d.f.s*. The biggest gap, often represented by  $D$ , has a known asymptotic distribution that is used to convert the test into the KS-test  $p$  value, but we do not need this additional step. We simply use  $D$  as the metric of fit; as the  $D$  value increases, less of the observed  $p$ -curve is captured by the expected  $p$ -curve.  $D$  has an intuitive representation, if  $D = .4$ , then the biggest observed gap in (cumulative)  $p$ -curve is 40%. For example, 78% of  $p$  values are supposed to be  $p < .041$ , but only 38% of them are:  $78\% - 38\% = 40\%$ . In Supplement 4, we discuss alternatives to the KS test for measuring fit.

#### A3. Minimizing the loss

The last step consists of finding the candidate effect size that minimizes  $D$ . We rely on R's *optimize()* command for this, but we first exhaustively search the plausible space of effect size so as to (a) reduce the odds that *optimize()* lands on a local rather than global minimum and (b) provide a diagnostic plot of how well different effect size fit the observed  $p$ -curve. See R code below.

# R-CODE for estimating effect size via  $p$ -curve – written by Uri Simonsohn

# Define the loss function

```
loss=function(t_obs,df_obs,d_est) {
  t_obs=abs(t_obs)
  p_obs=2*(1-pt(t_obs,df=df_obs))
  t.sig=subset(t_obs,p_obs<.05)
  df.sig=subset(df_obs,p_obs<.05)
  ncp_est=sqrt((df.sig+2)/4)*d_est
  tc=qt(.975,df.sig)
  power_est=1-pt(tc,df.sig,ncp_est)
```

#Syntax t\_obs: vector of t-values, df\_obs of degrees of freedom, d\_est: candidate d

#Take absolute value of t-value ( $p$ -curve assumes same sign and/or sign does not matter)

#Compute  $p$ -values of each  $t$  in  $t\_obs$  so as to keep only  $p < .05$  results

#Significant t-values

#d.f. associated with significant t-values

#Compute noncentrality parameter for that sample size and candidate effect size

#Compute critical t-value to get  $p = .05$

#Compute power for obtaining that t-value or bigger, given the noncentrality parameter



```

p_larger=pt(t.sig,df=df.sig,ncp=ncp_est) #Probability of obtaining a t-value bigger than the one that is observed (this is a vector)
ppr=(p_larger-(1-power_est))/power_est #Conditional probability of larger t-value given that it is  $p < .05$ ,  $pp$ -values
KSD=ks.test(ppr,punif)$statistic #Kolmogorov Smirnov test on that vector against the theoretical U[0,1] distribution
return(KSD) }

#Find the best fitting effect size (this also generates a diagnostic plot)

plotloss=function(t_obs,df_obs,dmin,dmax) #Syntax, same as above plus: dmin/dmax: smallest/biggest d considered,
{ loss.all=c() #Vector where results of fit for each candidate effect size are stored
  di=c() #Vector where the respective effect sizes are stored
  for (i in 0:((dmax-dmin)*100)) { #Do a loop considering every effect size between dmin and dmax in steps of .01
    d=dmin+i/100 #What effect size are we considering?
    di=c(di,d) #Add it to the vector of effect sizes
    options(warn=-1) #turn off warning because R often generates warnings when using noncentral pt() and
    qt() that are inconsequential (they involve lack of precision at a degree where precision lacks practical relevance)
    loss.all=c(loss.all,loss(df_obs=df_obs,t_obs=t_obs,d_est=d)) #add loss for that effect size to the vector with all losses
    options(warn=0) #turn warnings back on
  }
  imin=match(min(loss.all),loss.all) #Find the attempted effect size that leads to smallest loss overall
  dstart=dmin+imin/100 #Counting from dmin, what effect size is that?
  dhat=optimize(loss,c(dstart-.1,
    dstart+.1), df_obs=df_obs,t_obs=t_obs) #Now optimize in the neighborhood of that effect size

#PLOT RESULTS

plot(di,loss.all,xlab="Effect size\nCohen-d", ylab="Loss (D stat in KS test)",ylim=c(0,1), main="How well does each effect size fit?
(lower is better)")
points(dhat$minimum,dhat$objective,pch=19,col="red",cex=2) #Put a red dot in the estimated effect size
#Add a label
text(dhat$minimum,dhat$objective-.08,paste0("p-curve's estimate of effect size:\nd=",round(dhat$minimum,3)),col="red")
return(dhat$minimum)
}

#Example
t_obs= c(1.7, 2.8, -3.1, 2.4) # include one  $p > .05$  and one negative t-value to highlight how we treat those
df_obs=c(44, 75, 125, 200)
plotloss(t_obs=t_obs,df_obs=df_obs,dmin=-1,dmax=1)

```

## Acknowledgments

We received useful feedback for this project at the following seminar series and conferences: Wharton Decision Processes (March, 2012), Harvard NOM (April 2012), Columbia Marketing (May 2012), UCLA Marketing (May, 2012), BEAM – Cornell (June, 2012), Columbia Statistics (September 2012), Rutgers Psychology (October, 2012), Berkeley Initiative for Transparency in Social Science (December, 2012), UCSD Rady School (January, 2013), SPSP conference, New Orleans (January 2013), University of Southern California Marketing (February, 2013), the Solid Psychological Science Symposium, Nijmegen (June, 2013), and INSEAD (April, 2014). Any errors are our responsibility. This manuscript previously circulated as: Nelson, Simonsohn, & Simmons “P-Curve Fixes Publication Bias: Obtaining Unbiased Effect Size Estimates from Published Studies Alone.”

## Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

## Supplemental Material

Additional supporting information may be found at <http://pps.sagepub.com/content/by/supplemental-data> or at [http://www.p-curve.com/Supplement/Supplement\\_pcurve2.pdf](http://www.p-curve.com/Supplement/Supplement_pcurve2.pdf)

## Notes

1. The two sample  $t$  test with 38 degrees of freedom obtains  $p = .05$  if  $t(38) = 2.024$ . We can find the corresponding effect size for  $t = 2.024$  by recalling that  $t = \frac{M_2 - M_1}{SD \sqrt{2/n}}$  and  $\hat{d} = \frac{M_2 - M_1}{SD}$ , so  $\hat{d} = t \sqrt{2/n}$ . Needing  $t \geq 2.024$  for statistical significance is hence mathematically equivalent to needing  $\hat{d} \geq 2.024 \sqrt{2/20} = .64$ .
2. See the User Guide section for more details regarding what  $p$  values may and may not be included in  $p$ -curve.
3. For ease of exposition, throughout the article we focus on differences of means  $t$  tests and, hence, the  $t$  distribution.
4. Our assessment of Trim and Fill being the most commonly used corrective technique is informed by two sources. First, our casual observation indicates that, in psychology, review articles that correct for publication bias do so exclusively using Trim and Fill. Second, as of May 2014, Google Scholar indexes the

original Trim and Fill article as having 1,504 citations. Articles introducing other correction tools reviewed by Hedges and Vevea (2005) have about 50–150 citations. Thus, a rough estimate is that Trim and Fill is at least 10 times as popular as its competitors.

5. With this method, a meta-analyst explores the relation between the sample size and effect size of a set of studies, looking to see whether some studies appear to be missing as a result of publication bias. For example, if a very large-sample study estimates an effect size to be  $\hat{d} = .40$ , and most of the smaller studies estimate an effect to be greater than  $\hat{d} = .40$ , then it is presumed that smaller studies estimating effects less than  $\hat{d} = .40$  were unpublished. Trim and Fill is an algorithm that trims (i.e., eliminates) some real studies, and fills in (i.e., introduces) some non-real ones, seeking to obtain a final set of studies in which there would be a similar number of small sample studies above and below  $\hat{d} = .40$ .

6. Duval and Tweedie (2000a) write “Our key assumption is that the suppression has taken place in such a way that the [...] most extreme negative values have been suppressed” (p. 91).

7. Hedges and Vevea (2005) reviewed additional publication bias correction tools that, as we do here, assume selective reporting based on  $p$  values. The approach most similar to ours is by Hedges (1984). Two key differences are that his approach does not eliminate bias due to selective reporting of studies (see his Fig. 4), and cannot be applied when all effects are of the same sign (Hedges & Vevea, 2005, p. 152).

8. To provide a more intuitive treatment of heterogeneity than that captured in Figure 2C’s simulations, we performed additional simulations where half the studies had one true underlying effect size ( $d_1$ ), and the other half had a different true underlying effect size ( $d_2$ ). We then applied  $p$ -curve to that pooled set of statistically significant findings and verified that the estimated effect size,  $\hat{d}$ , corresponded to the average of the two true effect sizes. For example, if we set  $d_1 = d_2 = .4$ , then  $p$ -curve estimates  $\hat{d} = .4$ . But  $p$ -curve also estimates  $\hat{d} = .4$  if we set  $d_1 = .3$  and  $d_2 = .5$ , or  $d_1 = .2$  and  $d_2 = .6$ , such that the average true effect is  $.4$ .  $p$ -curve is robust to heterogeneity in effect size across studies (see Supplement 2 for more details and additional variations).

9. Readers wanting even more details can see the R code behind Figure 3 in Supplement 4.

10. Note that we already showed  $p$ -curve to be robust to heterogeneity in effect size. We focus on homogenous studies not to ensure that  $p$ -curve is valid, but to ensure our benchmark for truth is.

11. Their  $I^2$  statistic across labs, which measures the percent of variance explained by lab heterogeneity (Higgins, Thompson, Deeks, & Altman, 2003), was <10% for sunk costs, and <.01% for Asian disease. Neither of the effects were associated with significant differences between American and non-American labs, nor between laboratory and online methods of data collection (all  $ps > .29$ ; see Table 3 in Klein et al., 2014).

12. The sunk cost problem consisted of a question asking participants to rate their willingness to attend a match of their favorite team on a freezing cold day either if they had paid for a ticket or if it was free (for earlier studies of this effect, obtaining bigger effects probably due to more precise wording, see Arkes & Blumer, 1985; Thaler, 1980). The Asian disease

problem consisted of a binary hypothetical choice problem that presented participants with the same information using either a gains or a losses frame (Tversky & Kahneman, 1981).

13. Two of the 36 labs had disproportionately large sample sizes,  $N > 1,000$ . To make things more interesting, the results reported in the main text exclude those two labs so that there is more variability (we use 34 different labs). Results with those two labs are reported in the next footnote.

14. Figure 5 was constructed excluding the two largest labs. With them included, the effect size estimates for the sunk cost fallacy are  $\hat{d}_{\text{naive}} = .33$ ,  $\hat{d}_{\text{Earnest}} = .28$ ,  $\hat{d}_{p\text{-curve}} = .28$ . For the Asian Disease these are  $\hat{d}_{\text{naive}} = .63$ ,  $\hat{d}_{\text{Earnest}} = .60$ ,  $\hat{d}_{p\text{-curve}} = .60$ . We deemed these results less interesting because the very large samples greatly reduce bias in the subset of  $p < .05$  studies.

15. The average test statistic is  $\frac{2.12+2.43}{2} = 2.28$ , leading to  $t(38) = 2.28$ ,  $p = .028$ .

16. The reason not to include average  $p$  values nor average test statistics is that they are not uniform under the null. For instance, as one averages more and more  $p$  values the result converges to .5 (rather than to a uniform distribution of 0 to 1).

17. It is easy to specify  $p$ -curve in a way that allows for effects of opposite sign to be included simultaneously, but we decided against it. The reason is that it is quite unlikely that a true underlying effect leads to significant results of opposite sign. The probability of getting a  $p < .05$  effect of the “wrong” sign is necessarily smaller than 2.5% (that’s the probability if  $d = 0$ ). If a study is powered to just 20%, the odds are less than 1 in 1,000, and if a study is powered to 50%, the odds are less than 1 in 11,000. If statistically significant opposite sign effects are observed, separate analyses for those  $d > 0$  and  $d < 0$  are almost surely more meaningful and informative. Our choice-overload demonstration exemplifies how splitting effects by sign can lead to insightful inferences.

18. In the Appendix, we explain the procedure and provide R Code for a difference of means  $t$  test. Because the mapping between a test results (e.g.,  $t$  value) and effect size (e.g., Cohen’s  $d$ ) is different for each design, users of  $p$ -curve will need to either create custom programs for different designs (e.g., interactions, regressions, mixed designs), or more conveniently, use  $p$ -curve to estimate average power and then convert the average power to a measure of effect size. Supplement 4 includes the R program that computes average power.

19. Recall that we assume all  $t$  values are of the same sign, so we do not count significant effects of the wrong sign into power calculations.

20. Differences across  $p$ -curves for the same level of power depend on the degrees of freedom of the  $t$  test (on how thick the tails of the  $t$  distribution are), and converge to the normal distribution’s  $p$ -curve as the degrees of freedom increase.

## References

- Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, 35, 124–140.
- Ashenfelter, O., Harmon, C., & Oosterbeek, H. (1999). A review of estimates of the schooling/earnings relationship, with tests for publication bias. *Labour Economics*, 6, 453–470.



- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376.
- Chapman, G., & Johnson, E. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes*, *79*, 115–153.
- Chernev, A., Bockenholt, U., & Goodman, J. (2010). Commentary on Scheibehenne, Greifeneder, and Todd choice overload: Is there anything to it? *Journal of Consumer Research*, *37*, 426–428. doi:10.1086/655200
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.
- Cole, L. C. (1957). Biological clock in the unicorn. *Science*, *125*, 874–876.
- Cumming, G. (2008). Replication and P intervals: P values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286–300.
- Duval, S., & Tweedie, R. (2000a). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89–98.
- Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463.
- Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.
- Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician*, *60*, 328–331.
- Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research do arbitrary significance levels distort published results? *Sociological Methods & Research*, *37*, 3–30.
- Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational and Behavioral Statistics*, *9*, 61–85.
- Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis* (pp. 145–174). Chichester, England: Wiley.
- Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557.
- Hung, H. M. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The behavior of the P-value when the alternative hypothesis is true. *Biometrics*, *53*, 11–22.
- Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–646.
- John, L., Loewenstein, G. F., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*, 524–532.
- Klein, R. A., Ratliff, K., Vianello, M., Reginald, B., Adams, J., Bahník, S., . . . Nosek, B. A. (2014). *Investigating variation in replicability: A “Many Labs” Replication Project*. Retrieved from Open Science Framework osf.io/wx7ck
- Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*, 107–112.
- Oppenheimer, D., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.
- Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical psychology*, *58*, 646–656.
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis*. West Sussex, England: Wiley.
- Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research*, *37*, 409–425.
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Dialogue: The Official Newsletter of the Society for Personality and Social Psychology*, *26*(2), 4–7.
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*, 30–34.
- Sterling, T. D., Rosenbaum, W., & Weinkam, J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician*, *49*, 108–112.
- Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, *1*, 39–60.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.
- Wallis, W. A. (1942). Compounding probabilities from independent significance tests. *Econometrica*, *10*, 229–248.

# Exhibit 62

# Information Uncertainty and Stock Returns

X. FRANK ZHANG\*

## ABSTRACT

There is substantial evidence of short-term stock price continuation, which the prior literature often attributes to investor behavioral biases such as underreaction to new information. This paper investigates the role of information uncertainty in price continuation anomalies and cross-sectional variations in stock returns. If short-term price continuation is due to investor behavioral biases, we should observe greater price drift when there is greater information uncertainty. As a result, greater information uncertainty should produce relatively higher expected returns following good news and relatively lower expected returns following bad news. My evidence supports this hypothesis.

THERE IS SUBSTANTIAL EVIDENCE OF SHORT-TERM stock price continuation, which the prior literature often attributes to investor underreaction to new information. Examples include the positive serial correlation of returns at 3- to 12-month horizons (Jegadeesh and Titman (1993)), post-earnings announcement stock price drift in the direction indicated by the earnings surprise, and post-event return drift in the direction of the announcement date return.<sup>1</sup>

In this paper I investigate how information uncertainty contributes to this phenomenon. By information uncertainty, I mean ambiguity with respect to the implications of new information for a firm's value, which potentially stems from two sources: the volatility of a firm's underlying fundamentals and poor information.<sup>2</sup> My main hypothesis is that if investors underreact to public

\*X. Frank Zhang is from Yale University. I am grateful to Anwer Ahmed, Ray Ball, Daniel Bens, Philip Berger, Kent Daniel, Rachel Hayes, Charles Lee, Richard Leftwich, Joseph Piotroski, Robert Stambaugh (the editor), Richard Thaler, Franco Wong, and participants at the University of Chicago Asset Pricing Lunch session for many helpful comments, and especially Robert Bushman, Stephanie Curcuru, and Abbie Smith for extensive discussion and editorial assistance. Special thanks also go to an anonymous referee for many constructive suggestions. Chris Malloy provided me with the unadjusted individual forecast data under permission from I/B/E/S. Any remaining errors or ambiguities are solely my responsibility.

<sup>1</sup> See the appendix of Daniel, Hirshleifer, and Subrahmanyam (1998) for a thorough review of this evidence.

<sup>2</sup> Theoretically, an observed signal ( $s$ ) is characterized as a firm's fundamental value ( $v$ ), such as future cash flow or dividend, plus a noise term ( $e$ ), that is,  $s = v + e$ . The variance of the signal measures information uncertainty:  $var(s) = var(v) + var(e)$ , where  $var(v)$  is a firm's underlying fundamental volatility and  $var(e)$  reflects the quality of information. I do not distinguish a firm's underlying fundamental volatility from information quality because both effects contribute to the uncertainty of a firm's value and because it is hard to empirically disentangle one from the other as observed stock volatility and other empirical constructs capture both effects. This definition parallels the argument in Hirshleifer (2001).

information, they will underreact even more in cases of greater information uncertainty. The testable implication is that greater information uncertainty about the impact of news on stock value leads to higher expected stock returns following good news but lower expected stock returns following bad news relative to the returns of stocks about which there is less information uncertainty. A distinct feature of the analysis is the focus on how price continuation following the release of public information varies with information uncertainty.

My hypothesis is motivated by two results from the behavioral finance literature. Several papers including Chan, Jegadeesh, and Lakonishok (1996) attribute price continuation to a gradual market response to information. Hirshleifer (2001) and Daniel, Hirshleifer, and Subrahmanyam (1998, 2001) posit that psychological biases are increased when there is more uncertainty. This study combines these two ideas and tests the following joint hypothesis: If the slow market response to information is due to psychological biases such as overconfidence, these psychological biases will be larger and, hence, the price response will be slower when there is more ambiguity about the implications of the information for a firm's value.

Specifically, I study two price continuation anomalies: post-analyst forecast revision price drift and price momentum. I focus on these two anomalies because the new information is public, easily categorized as good or bad, and occurs fairly frequently. For the first anomaly, the new information is the current month's earnings forecast revision. For the second, the new information is the average monthly stock returns over the past 11 months. I classify upward forecast revisions or past winners as good news and downward revisions or past losers as bad news.

Using ex post returns as a proxy for expected returns, I find consistent results across six proxies for information uncertainty: firm size, firm age, analyst coverage, dispersion in analyst forecasts, return volatility, and cash flow volatility.<sup>3</sup> For each of the six proxies, I find that greater information uncertainty leads to relatively lower future stock returns following bad news and relatively higher future returns following good news, suggesting that uncertainty delays the flow of information into stock prices.<sup>4</sup> In other words, the market reaction to

<sup>3</sup> In all analyses, I construct the proxies in such a way that a higher value corresponds to greater information uncertainty. Specifically, I use the reciprocals of firm size, firm age, and analyst coverage. To ensure that all information is available before the portfolio formation date, I use all sorting variables as of the current month and predict 1-month-ahead stock returns.

<sup>4</sup> The paper studies both the mean and interaction effects of information uncertainty. By the mean effect, I refer to the effect of information uncertainty on future stock returns unconditional on the nature of news. I investigate whether high-uncertainty stocks earn relatively higher future returns than low-uncertainty stocks. By the interaction effect, I mean the interaction between information uncertainty and the nature of news. I predict that high-uncertainty stocks have relatively higher (lower) future returns than low-uncertainty stocks following good (bad) news, that is,  $RET_H^G > RET_L^G$  and  $RET_H^B < RET_L^B$ , where  $RET_H^G$  and  $RET_L^G$  ( $RET_H^B$  and  $RET_L^B$ ) are returns for high- and low-uncertainty stocks following good (bad) news, respectively. I also study how the performance of certain trading strategies varies with information uncertainty. I predict that the momentum strategy works better for high-uncertainty stocks, that is,  $RET_H^G - RET_H^B > RET_L^G - RET_L^B$ . It can easily be shown that the effect of information uncertainty conditional on the nature of news, as hypothesized in the paper, is a sufficient but not a necessary condition for the stronger momentum effect for high-uncertainty stocks.

new information is relatively complete for low-uncertainty stocks, and there is little news-based return predictability. For high-uncertainty stocks, on the other hand, the market reaction is far from complete. Good news predicts relatively higher future returns and bad news predicts relatively lower future returns. This relation between information uncertainty and future returns remains after I control for common factors used in prior empirical studies. I provide further assurance that missing risk factors do not drive the results by documenting a similar return pattern around subsequent earnings announcement dates.

The opposite effects of information uncertainty on stock returns following good versus bad news amplify the results of previously documented trading strategies. As a result, trading strategies that buy good-news stocks and short bad-news stocks work particularly well when limited to high-uncertainty stocks.<sup>5</sup> For example, a momentum strategy (buying past winners and shorting past losers) on stocks in the bottom stock volatility quintile (low uncertainty) generates a 0.63% monthly return, but a similar strategy based on stocks in the top stock volatility quintile (high uncertainty) yields a 2.63% monthly return. Other uncertainty proxies produce similar returns.

My main prediction is related to the theoretical work of Daniel et al. (1998, 2001), but has broader implications.<sup>6</sup> Daniel et al. (1998) develop a model in which investors are overconfident about their private information, and therefore overweight their private information and underreact to public signals (e.g., analyst forecast revisions). As a result, future returns are predictable. Daniel et al. (1998, 2001) further argue that the return predictability should be stronger in firms with greater uncertainty because investors tend to be more overconfident when firms' businesses are hard to value. This argument implies that greater uncertainty is related to relatively higher (lower) stock returns following good (bad) news. Because I do not incorporate measures of private information or overconfidence in my empirical analysis, my evidence leaves the door open for other behavioral models. For example, my results are also consistent with a behavioral model in which investors overweight their priors relative to new information due to the anchoring/conservatism bias and overweight their priors more when there is greater information uncertainty.

This study contributes to the accounting and finance literature in several ways. First, the paper provides evidence in support of the hypothesis that price continuation following public signals increases with proxies for the ambiguity of the signals with respect to the implications for a firm's value. The fact that proxies for information uncertainty, such as cash flow and stock return volatility, are associated with both higher returns following good news and lower returns following bad news but are not significantly related to unconditional expected returns suggests that momentum effects are more likely to

<sup>5</sup> Prior literature finds that the momentum strategy works better for small firms, growth firms, firms with low analyst following, and firms with high abnormal trading volume (see Daniel and Titman (1999), Hong, Lim, and Stein (2000) Lee and Swaminathan (2000)). Such evidence is in general accordance with my prediction on the interaction effect between information uncertainty and momentum.

<sup>6</sup> Also see Jiang, Lee, and Zhang (2004) for the argument supporting the position that the information uncertainty effect is associated with investor overconfidence and arbitrage costs.

reflect slow absorption of ambiguous information into stock price than to reflect missing risk factors.

Second, the evidence presented here sheds new light on the role of accounting disclosure in capital market settings. In the prior literature, information uncertainty is often modeled as the information asymmetry component of the cost of capital (e.g., Diamond and Verrecchia (1991), Easley and O'Hara (2001), Verrecchia (2001)) or estimation risk (e.g., Barry and Brown (1985), Coles and Loewenstein (1988), Klein and Bawa (1976)) and therefore increases expected stock returns. The theoretical argument that accounting disclosure can reduce information uncertainty and cost of capital is appealing, but the overall empirical evidence is mixed.<sup>7</sup> My evidence that the effects of information uncertainty on future returns following good and bad news offset each other in unsigned analysis might explain why previous studies often find an insignificant effect of accounting disclosure (see the review by Verrecchia (2001)). My evidence also suggests a potential additional role for accounting disclosure: More transparent disclosure might reduce information uncertainty and speed the absorption of new information into the stock prices.

Finally, the evidence also questions the underlying cause of the size effect. The prior literature finds that small stocks historically earned higher returns than large stocks, but that this effect has disappeared in the last 20 years. As shown in this paper, the opposite effects of size on stock returns following good and bad news suggest that firm size behaves more like a proxy for information uncertainty than a common risk factor in the cross section of stock returns. The positive (negative) size premium following good (bad) news is also persistent over time. These results also have implications for studies using firm size or other variables related to information uncertainty as control variables in unsigned cross-section analysis.

The remainder of the paper is organized as follows. The next section discusses related literature and outlines my main prediction. Section II describes the sample data and provides descriptive statistics. Section III examines the role of information uncertainty from a portfolio approach. Section IV uses a four-factor model to control for some common factors. Section V examines stock price reactions to earnings announcements following the portfolio formation date. Section VI conducts some robustness checks, and Section VII concludes.

## **I. Related Literature and Hypothesis Development**

There is substantial evidence of short-term stock price continuation. For example, Stickel (1991), Chan et al. (1996), and Gleason and Lee (2003) document that stock prices exhibit a drift after analyst forecast revisions. Forming

<sup>7</sup> For example, Botosan (1997) finds a negative association between a self-constructed disclosure index and the cost of capital but only for firms followed by few analysts. Using AIMR scores as a proxy for disclosure level, Botosan and Plumlee (2002) find that the cost of capital is negatively related to annual report disclosure level but positively related to quarterly report disclosure level. Finally, Cohen (2003) reports that the negative relation between disclosure and the cost of capital disappears once he controls for the endogeneity associated with the reporting quality choice.



portfolios based on past intermediate-horizon stock returns, Jegadeesh and Titman (1993) show that past winners on average continue to outperform past losers over the next 3 to 12 months. Several papers including Bernard and Thomas (1990) demonstrate that stock prices continue to drift in the direction of quarterly earnings surprises for at least 120 trading days following the earnings announcement.

Short-term stock price continuation is often attributed to investor behavioral biases such as investor underreaction to new information. Chan et al. (1996) show that the post-analyst revision drift is part of a general class of “momentum” strategies, in which the market response to recently released information is gradual so that prices exhibit predictable drift patterns. Chan et al. (1996) and Barberis, Shleifer, and Vishny (1998), among others, argue that the intermediate-horizon price momentum effect is due to investor underreaction to some information.<sup>8</sup> Daniel et al. (1998) develop a model in which investors are overconfident with their private information and therefore underreact to public signals. This model provides a potential explanation for the underlying cause of post-analyst revision drift or momentum.<sup>9</sup>

Hirshleifer (2001) posits that greater uncertainty about a set of stocks and a lack of accurate feedback about their fundamentals leave more room for psychological biases. Therefore, the misvaluation effects of almost any mistaken-beliefs model should be strongest among firms about which there is high uncertainty and poor information. For example, Daniel et al. (1998, 2001) show that return predictability should be stronger in firms with greater uncertainty because investors tend to be more overconfident when firms’ businesses are hard to value.

I combine these two ideas and test the following joint hypothesis: If post-analyst revision drift, momentum, and other short-term anomalies are due to investor psychological biases such as overconfidence, we should observe greater investor behavioral biases and stronger price drifts when there is greater information uncertainty. The testable implication is that greater information uncertainty produces relatively higher (lower) stock returns following good (bad) news. The opposite effects of information uncertainty on future stock returns following good and bad news also amplify the profitability of certain trading strategies. As a result, the momentum trading strategy works particularly well when limited to high-uncertainty stocks.

I use two measures of news. First, I use analyst forecast revisions for the current month. An upward revision means good news, and a downward revision means bad news. Although this measure may be noisy since analysts may suffer from behavioral biases or have incentives to bias their forecasts,

<sup>8</sup> It seems safe to classify momentum as an underreaction story in my setting, as I focus on a short return window (1 month) and Lee and Swaminathan (2000) and Jegadeesh and Titman (2001) find that the price momentum effect only partially reverses over long horizons (5 years).

<sup>9</sup> While momentum in the correction phase and virtually all postevent price drifts are classified as investor underreaction to public signals in Daniel et al. (1998), they offer a different mechanism for momentum in the overreaction phase. Namely, investors keep overreacting to their priors because of biased self-attribution, which contributes positively to short-term momentum.



it is relevant as long as analysts on average react in the same direction as the news suggests. Measurement error in my variables works against finding any significant results. My second measure is past stock returns. If investors follow the direction of new information, a partition based on price momentum (the past 11-month stock returns) is another way to distinguish good news from bad.

I also need a proxy for information uncertainty. One natural variable is firm size (MV), measured as the market capitalization at the portfolio formation date. It seems plausible that small firms are less diversified and have less information available for the market than large firms. Small firms may also have fewer customers, suppliers, and shareholders, and may not bear high disclosure preparation costs. Investors might have fixed costs of information acquisition, which makes small firms' stocks unattractive. Unfortunately, even if firm size is, in fact, a useful measure of uncertainty, it is likely to capture other things as well, potentially confounding any inferences. I therefore use five alternative proxies for information uncertainty: firm age, analyst coverage, dispersion in analyst forecasts, stock volatility, and cash flow volatility. Although each proxy might also capture other effects, the common element is their ability to quantify information uncertainty.

Firms with a long history have more information available to the market (Barry and Brown (1985)). To the extent that older firms are more likely to be in more mature industries, firm age also captures the underlying volatility at the industry level. I use firm age (AGE) as my second proxy, measured as the number of years since the firm was first covered by the Center for Research in Securities Prices (CRSP). To my knowledge, the role of firm age in predicting future returns has not been empirically documented in the prior literature.

A third proxy is analyst coverage (COV), measured as the number of analysts following the firm in the previous year. Analysts collect, digest, and distribute information about a firm's performance. There is evidence that larger analyst coverage is likely to correspond to more information available about the firm, which implies less uncertainty. Lang and Lundholm (1996) find that analyst coverage is positively associated with disclosure scores. Hong, Lim, and Stein (2000) use larger analyst coverage as an indicator of less information asymmetry. Gleason and Lee (2003) show that the post-revision price drift is more pronounced in firms with smaller analyst coverage.

The fourth proxy is dispersion in analyst earnings forecasts (DISP). In the prior literature, forecast dispersion is widely used to proxy for the uncertainty about future earnings or the degree of consensus among analysts or market participants (e.g., Barron et al. (1998), Barron and Stuerke (1998), Diether, Malloy, and Scherbina (2002), Imhoff and Lobo (1992), Lang and Lundholm (1996)). I measure forecast dispersion as the standard deviation of analyst forecasts scaled by the prior year-end stock price to mitigate heteroskedasticity.

The fifth proxy is stock volatility (SIGMA), which is measured by the standard deviation of weekly market excess returns over the year ending at the portfolio formation date. Following Lim (2001), I measure weekly returns from Thursday to Wednesday to mitigate nonsynchronous trading or bid-ask bounce effects

in daily prices. A 1-year estimation period is chosen to provide a reasonable number of observations.

The final proxy is cash flow volatility (CVOL), measured as the standard deviation of cash flow from operations in the past 5 years (with a minimum of 3 years). I treat CVOL as missing if there are only 1 or 2 years' data available. Cash flow from operations is earnings before extraordinary items (Compustat #18) minus total accruals, scaled by average total assets (Compustat #6), where total accruals are equal to changes in current assets (Compustat #4) minus changes in cash (Compustat #1), changes in current liabilities (Compustat #5), and depreciation expense (Compustat #14) plus changes in short-term debt (Compustat #34).<sup>10</sup> Although the cash flow measure is indirectly calculated from financial statements and therefore is affected by a firm's information system, it is more likely to capture the underlying volatility.

## **II. Sample Data and Descriptive Statistics**

The sample data come from three sources. Returns are from the CRSP Monthly Stocks Combine File, which includes NYSE, AMEX, and Nasdaq stocks. Book value and other financial data are from Compustat. Analyst forecast revisions are from I/B/E/S.<sup>11</sup> The sample period spans from January 1983 to December 2001.

I delete observations for which the absolute value of earnings forecast revision exceeds 100% of the prior year-end stock price, because these observations are likely to be erroneous. Following Jegadeesh and Titman (2001), I exclude stocks with a share price below \$5 at the portfolio formation date to make sure that the results are not driven by small, illiquid stocks or by the bid-ask bounce. To avoid any potential confounding effect of recent IPOs, I also exclude firms with less than 12 months of past return data on CRSP.

Table I presents descriptive statistics for variables of interest. The mean monthly return is 1.15% and the median is 0.74%, indicating a slight right skewness in the distribution. Although I/B/E/S tends to cover large firms, there is a large variation in firm size in my sample. The market value ranges from

<sup>10</sup> This balance sheet approach to estimate accruals may be subject to the measurement error problem (see Hribar and Collins (2002)), but it is unavoidable as cash flow statements are not available until 1987. The results are robust in the post-1987 period using accruals from the cash flow approach.

<sup>11</sup> There are two problems with the standard-issue I/B/E/S summary data set. First, I/B/E/S uses all existing analyst forecasts to calculate summary statistics, and some of these forecasts are stale. These stale forecasts tend to increase the dispersion in analyst forecasts. Second, there is a rounding error problem with stock splits because I/B/E/S adjusts all data for stock splits and only rounds the estimate to the nearest cent (Baber and Kang (2002)). For example, the adjustment factor for Dell during the 1988 to 1991 period is 96, which renders virtually all forecast revisions to be zero in the I/B/E/S-adjusted Detail History File. Dell had \$1.35 actual earnings per share in 1990. Any forecast between \$0.48 and \$1.44 would have the same adjusted \$0.02 earnings per share in the I/B/E/S database. The rounding error problem tends to reduce both forecast revisions and forecast dispersion. To avoid these issues, I follow Diether et al. (2002) and Zhang (2005) and calculate forecast revisions and other variables based on the raw detail forecast data unadjusted for stock splits. However, the results are robust to the standard-issue I/B/E/S data set.

Table I  
Descriptive Statistics

Firm size (MV) is the market capitalization (in millions of dollars) at the end of month  $t$ . Book-to-market (BM) is the book value of equity divided by its market value at the end of the last fiscal year.  $RET_{t-11,t-1}$  is accumulated returns from months  $t - 11$  to  $t - 1$ . Firm age (AGE) is the number of years since the firm was first covered by CRSP. Analyst coverage (COV) is the number of analysts following the firm in the previous year. Forecast dispersion (DISP) is the standard deviation of analyst forecasts in month  $t$  scaled by the prior year-end stock price. Stock volatility (SIGMA) is the standard deviation of weekly market excess returns over the year ending at the end of month  $t$ . Cash flow volatility (CVOL) is the standard deviation of cash flow from operations in the past 5 years (with a minimum of 3 years), where cash flow from operations is earnings before extraordinary items minus total accruals estimated from the balance sheet approach, scaled by average total assets. Stocks with a price less than \$5 are excluded from the sample. The sample period is from January 1983 to December 2001.

Panel A: Descriptive Statistics								
	<i>N</i>	Mean	Std. Dev.	Min	Q1	Median	Q3	Max
$RET_{t+1}$	490,396	1.15%	13.72%	−98.13%	−5.56%	0.74%	7.32%	556%
$MV_t$	490,396	2,378	11,033	0	130	382	1,286	602,433
$BM_t$	490,396	0.642	0.462	0.000	0.326	0.550	0.849	20.941
$RET_{t-11,t-1}$	483,213	22.65%	68.21%	−98.16%	−10.42%	12.49%	39.49%	4608%
$AGE_t$	490,396	18	16	1	6	13	24	77
$COV_t$	458,263	11	10	1	4	7	15	67
$DISP_t$	420,499	0.72%	2.67%	0.00%	0.12%	0.29%	0.71%	638%
$SIGMA_t$	487,675	5.54%	2.96%	1.03%	3.48%	4.82%	6.78%	82.47%
$CVOL_t$	351,417	0.074	0.094	0.001	0.031	0.053	0.089	6.940

Panel B. Correlation Matrix (Pearson Correlations Are Shown above the Diagonal with Spearman Below)									
	$RET_{t+1}$	$MV_t$	$BM_t$	$RET_{t-11,t-1}$	$AGE_t$	$COV_t$	$DISP_t$	$SIGMA_t$	$CVOL_t$
$RET_{t+1}$	1	−0.006	0.014	0.017	0.005	0.005	−0.002	−0.026	−0.008
$MV_t$	0.014	1	−0.096	0.028	0.232	0.350	−0.034	−0.081	−0.057
$BM_t$	0.025	−0.180	1	−0.069	0.163	−0.006	0.185	−0.174	−0.153
$RET_{t-11,t-1}$	0.025	0.147	−0.050	1	−0.063	−0.058	−0.009	0.160	0.080
$AGE_t$	0.026	0.409	0.252	0.003	1	0.428	0.008	−0.378	−0.209
$COV_t$	0.023	0.725	−0.025	−0.033	0.404	1	0.000	−0.252	−0.144
$DISP_t$	−0.006	−0.204	0.416	−0.076	0.070	0.045	1	0.011	0.042
$SIGMA_t$	−0.054	−0.332	−0.280	−0.087	−0.491	−0.293	0.033	1	0.359
$CVOL_t$	−0.030	−0.294	−0.239	−0.013	−0.349	−0.245	0.059	0.492	1

\$70,000 to \$602 billion. Firm age ranges from 1 to 77 years. Young firms account for a considerable portion of the sample, which is partly due to the fact that after 1973, CRSP includes Nasdaq firms. Stock returns are volatile, as suggested by a mean SIGMA of 5.54% per week and a median of 4.82% per week.

Panel B shows the correlation matrix. The Pearson (Spearman) correlation between returns and book-to-market is 0.014 (0.025), which confirms the value premium in univariate tests. The size effect is negative in the Pearson measure but positive in the Spearman measure. Firm size, firm age, and analyst coverage

are positively correlated with each other and negatively correlated with stock volatility and cash flow volatility, supporting the idea that these proxies capture the same phenomenon. One exception is analyst dispersion, which is negatively correlated with firm size but not with firm age and analyst coverage. Firm size is highly correlated with analyst coverage (Pearson = 0.35 and Spearman = 0.725), but the correlation between firm size and firm age is only moderate (Pearson = 0.232 and Spearman = 0.409). Stock volatility is highly correlated with cash flow volatility (Pearson = 0.359 and Spearman = 0.492) but not with dispersion in analyst forecasts or any other proxy for information uncertainty, suggesting that these proxies might capture different aspects of information uncertainty.

### **III. Portfolio Effects of Information Uncertainty**

In this section I assign stocks to portfolios based on the nature of news and the level of information uncertainty in order to draw conclusions about the average returns for these classes of stocks. This is a standard approach in asset pricing, which reduces the variability in returns.

#### *A. Portfolio Returns by Information Uncertainty Proxy*

The first set of empirical tests examines the cross-sectional variation in stock returns by information uncertainty level (the mean effect) and verifies the existence of the momentum effect and the post-analyst revision drift for my sample. In Table II, Panel A, each month I sort stocks into 10 deciles using a proxy for information uncertainty. I find that high-uncertainty stocks tend to have lower future returns than do low-uncertainty stocks. However, none of the trading strategies with a long position in high-uncertainty stocks and a short position in low-uncertainty stocks yields statistically negative returns.<sup>12</sup> The evidence of lower returns for high-uncertainty stocks than for low-uncertainty stocks does not support the notion that information uncertainty is a cross-sectional risk factor and compensated by higher stock returns.

The last column in Table II, Panel A verifies the existence of the momentum effect for my sample. I sort stocks based on past 11-month stock returns and find that past winners on average outperform past losers by 2.22% ( $t = 5.38$ )

<sup>12</sup> In a more recent, complementary study, Jiang, Lee, and Zhang (2004) find a significant negative mean effect in a similar setting. The insignificant mean effect here might be partly due to my choice of a 1-month holding period. Because the literature usually measures the monthly return for a  $K$ -month holding period as the simple average of portfolio returns from strategies implemented in the current month and the previous  $K - 1$  months, the monthly return tends to be less volatile when  $K$  is larger. I focus on the 1-month holding period in order to pick up the strong information uncertainty effect in the first month following public signals, as the effect of information uncertainty quickly goes away following good news (see footnote 14 and Figure 1). My short sample period might also play a role. When I test the mean effect of information uncertainty using the expanded 1964 to 2003 sample period, I find that the size effect is significantly negative, the stock volatility effect is marginally negative, and the effect of firm age is insignificant.

Table II  
Portfolio Returns by Information Uncertainty Proxy, Past Returns,  
and Analyst Forecast Revision

This table reports average monthly portfolio returns sorted by each information uncertainty proxy and verifies the existence of the momentum effect and the postrevision drift for my sample. In Panel A, each month I sort stocks into 10 deciles based on an information uncertainty proxy in month  $t$  or the past 11-month stock returns. In Panel B, I sort stocks into three news categories based on analyst forecast revisions in month  $t$ . Firm size (MV) is the market capitalization (in millions of dollars) at the end of month  $t$ . Firm age (AGE) is the number of years since the firm was first covered by CRSP. Analyst coverage (COV) is the number of analysts following the firm in the previous year. Forecast dispersion (DISP) is the standard deviation of analyst forecasts in month  $t$  scaled by the prior year-end stock price. Stock volatility (SIGMA) is the standard deviation of weekly market excess returns over the year ending at the end of month  $t$ . Cash flow volatility (CVOL) is the standard deviation of cash flow from operations in the past 5 years (with a minimum of 3 years), where cash flow from operations is earnings before extraordinary items minus total accruals, scaled by average total assets.  $1/MV$ ,  $1/AGE$ , and  $1/COV$  are the reciprocals of MV, AGE, and COV, respectively. Stocks with a price less than \$5 at the portfolio formation date are excluded from the sample. Stocks are held for 1 month, and portfolio returns are equally weighted. The sample period is from January 1983 to December 2001;  $t$ -statistics in parentheses are adjusted for autocorrelation.

Panel A: 10 Decile Returns Sorted by Information Uncertainty Level or Momentum							
	Sorted by 1/MV	Sorted by 1/AGE	Sorted by 1/COV	Sorted by DISP	Sorted by SIGMA	Sorted by CVOL	Sorted by $RET_{t-11,t-1}$
D1 (low)	1.18%	1.22%	1.18%	1.21%	1.40%	1.33%	0.13%
D2	1.28%	1.38%	1.20%	1.28%	1.44%	1.37%	0.87%
D3	1.21%	1.25%	1.18%	1.33%	1.40%	1.25%	1.03%
D4	1.23%	1.33%	1.19%	1.31%	1.39%	1.22%	1.15%
D5	1.29%	1.42%	1.22%	1.40%	1.35%	1.28%	1.18%
D6	1.30%	1.38%	1.17%	1.40%	1.29%	1.26%	1.34%
D7	1.42%	1.16%	1.12%	1.28%	1.34%	1.28%	1.47%
D8	1.40%	1.20%	1.08%	1.17%	1.24%	1.32%	1.47%
D9	1.23%	1.13%	1.26%	1.30%	1.05%	1.24%	1.76%
D10 (high)	1.15%	0.96%	1.09%	1.15%	0.72%	0.94%	2.35%
D10-D1	-0.02% (-0.10)	-0.26% (-0.61)	-0.09% (-0.66)	-0.06% (-0.25)	-0.68% (-1.04)	-0.38% (-0.91)	2.22% (5.38)

Panel B: Portfolio Returns Based on Analyst Forecast Revision			
	Sample	$RET_t$	$RET_{t+1}$
Bad news ( $REV_t < 0$ )	30.3	-0.20 (-0.53)	0.72 (1.99)
No news ( $REV_t = 0$ )	46.5%	1.92% (5.71)	1.29% (3.80)
Good news ( $REV_t > 0$ )	23.2%	3.55% (9.78)	1.84% (5.31)
Good – bad			1.13% (9.32)

in the first month after portfolio formation, which is consistent with the prior literature.

In Table II, Panel B, I sort stocks based on analyst forecast revisions to verify the existence of post-revision drift. An upward forecast revision means good news, and a downward revision means bad news. If a revision is zero, I assign it to a separate category. On average, negative, zero, and positive revisions account for 30.3%, 46.5%, and 23.2% of the sample data, respectively. Panel B shows that bad news corresponds to lower future returns and good news is followed by higher future returns, confirming the post-revision price drift documented in the prior literature. On average, bad-news firms gain 0.72% ( $t = 1.99$ ) in the following month, compared to 1.84% ( $t = 5.31$ ) for good-news firms. The return of 1.29% ( $t = 3.80$ ) for no-news firms falls in the middle. The returns for bad-, no-, and good-news firms are  $-0.20\%$ ,  $1.92\%$ , and  $3.55\%$ , respectively, in the month in which the revision news comes out. This pattern of future returns is consistent with the underreaction argument in the sense that investors underreact to new information and, as a result, future stock price movements are in the same direction as in the month in which the news occurs.

#### *B. Portfolio Returns by Analyst Forecast Revision and Information Uncertainty Proxy*

To test the relation between the nature of news and the effect of information uncertainty on future returns, in Table III I sort stocks by information uncertainty proxy for different news categories. Stocks are first classified into one of three categories based on their forecast revision in the current month. Within each revision category, stocks are sorted into 10 deciles by information uncertainty proxy. For the resulting 30 portfolios, there are an average of 66, 101, and 51 stocks each month for each portfolio in the bad-, no-, and good-news categories, respectively. To make sure that investors have all information available when forming portfolios, I use sorting variables as of the current month and predict 1-month-ahead stock returns.

Table III confirms my hypothesis. For each proxy, I observe that greater information uncertainty produces relatively lower future returns following bad news and relatively higher future returns following good news. For example, for the COV proxy, the mean portfolio return decreases from 0.94% in decile 1 (low uncertainty) to 0.10% in decile 10 (high uncertainty) in the bad-news category. A trading strategy with a long position in D10 stocks and a short position in D1 stocks ( $D10 - D1$ ) yields a  $-0.84\%$  ( $t = -2.96$ ) monthly return. For the good-news category, the mean portfolio return increases sharply from 1.63% in D1 to 2.28% in D10. The  $D10 - D1$  strategy yields a 0.65% monthly return ( $t = 2.27$ ).<sup>13</sup> For the no-news category, high-uncertainty stocks have slightly

<sup>13</sup> Unreported results show that analyst coverage for the bad-news category closely matches that of the good-news category, which excludes analyst coverage per se as a possible explanation for the difference in future returns for these two categories. Other uncertainty proxies in the bad-news category closely match those in the good-news category in each decile too.

Table III  
Portfolio Returns by Analyst Forecast Revision and Information  
Uncertainty Proxy

This table reports average monthly portfolio returns sorted by analyst forecast revision and information uncertainty proxy. Each month I sort stocks into three categories depending on whether the forecast revision is negative, zero, or positive. The forecast revision is the average of individual revisions by analysts who covered the firm in both months  $t - 1$  and  $t$ . For each category, I further sort stocks into 10 deciles based on information uncertainty proxy. Firm size (MV) is the market capitalization (in millions of dollars) at the end of month  $t$ . Firm age (AGE) is the number of years since the firm was first covered by CRSP. Analyst coverage (COV) is the number of analysts following the firm in the previous year. Forecast dispersion (DISP) is the standard deviation of analyst forecasts in month  $t$  scaled by the prior year-end stock price. Stock volatility (SIGMA) is the standard deviation of weekly market excess returns over the year ending at the end of month  $t$ . Cash flow volatility (CVOL) is the standard deviation of cash flow from operations in the past 5 years (with a minimum of 3 years), where cash flow from operations is earnings before extraordinary items minus total accruals, scaled by average total assets.  $1/MV$ ,  $1/AGE$ , and  $1/COV$  are the reciprocals of MV, AGE, and COV, respectively. Stocks with a price less than \$5 at the portfolio formation date are excluded from the sample. Stocks are held for 1 month, and portfolio returns are equally weighted. The sample period is from January 1983 to December 2001;  $t$ -statistics in parentheses are adjusted for autocorrelation.

	Sorted by 1/MV			Sorted by 1/AGE			Sorted by 1/COV		
	REV < 0	REV = 0	REV > 0	REV < 0	REV = 0	REV > 0	REV < 0	REV = 0	REV > 0
D1 (low)	1.00%	1.17%	1.41%	1.03%	1.13%	1.49%	0.94%	1.28%	1.63%
D2	0.82%	1.19%	1.60%	1.08%	1.26%	1.70%	0.97%	1.34%	1.64%
D3	1.04%	1.33%	1.69%	1.11%	1.56%	1.71%	0.99%	1.25%	1.71%
D4	0.96%	1.28%	1.60%	0.92%	1.33%	1.82%	0.78%	1.20%	1.59%
D5	0.82%	1.34%	1.62%	0.92%	1.65%	1.74%	0.74%	1.06%	1.72%
D6	0.72%	1.35%	1.83%	0.49%	1.34%	2.10%	0.58%	1.21%	2.08%
D7	0.61%	1.26%	2.02%	0.79%	1.00%	2.03%	0.69%	1.19%	1.83%
D8	0.55%	1.44%	2.13%	0.44%	1.18%	2.15%	0.56%	1.09%	1.93%
D9	0.30%	1.24%	2.52%	0.18%	1.21%	1.92%	0.31%	1.17%	2.05%
D10 (high)	0.12%	1.23%	2.36%	-0.16%	0.96%	1.97%	0.10%	1.10%	2.28%
D10 - D1	-0.87%	0.06%	0.96%	-1.19%	-0.17%	0.48%	-0.84%	-0.17%	0.65%
	(-3.02)	(0.23)	(2.88)	(-2.55)	(-0.38)	(1.03)	(-2.96)	(-1.05)	(2.27)

	Sorted by DISP			Sorted by SIGMA			Sorted by CVOL		
	REV < 0	REV = 0	REV > 0	REV < 0	REV = 0	REV > 0	REV < 0	REV = 0	REV > 0
D1 (low)	0.71%	1.29%	1.48%	1.25%	1.41%	1.71%	1.01%	1.25%	1.73%
D2	0.72%	1.26%	1.70%	1.12%	1.54%	1.73%	1.04%	1.42%	1.64%
D3	0.99%	1.34%	1.67%	1.04%	1.34%	1.81%	1.00%	1.19%	1.59%
D4	0.83%	1.28%	1.94%	0.91%	1.46%	1.75%	0.82%	1.24%	1.78%
D5	0.88%	1.38%	1.97%	0.79%	1.42%	1.77%	0.84%	1.44%	1.74%
D6	0.59%	1.44%	1.91%	0.63%	1.42%	1.76%	0.68%	1.18%	1.88%
D7	0.82%	1.31%	1.96%	0.54%	1.49%	1.94%	0.66%	1.43%	1.82%
D8	0.59%	1.20%	1.76%	0.53%	1.31%	1.90%	0.62%	1.19%	1.99%
D9	0.63%	1.28%	2.00%	0.37%	0.88%	2.24%	0.62%	1.00%	2.13%
D10 (high)	0.48%	1.08%	2.04%	-0.23%	0.51%	2.16%	0.05%	0.79%	2.27%
D10 - D1	-0.23%	-0.21%	0.56%	-1.47%	-0.90%	0.44%	-0.97%	-0.46%	0.54%
	(-0.77)	(-0.95)	(1.82)	(-2.23)	(-1.36)	(0.64)	(-2.10)	(-1.04)	(1.07)



lower returns than low-uncertainty stocks, but the difference is insignificant.<sup>14</sup> Other proxies for information uncertainty produce qualitatively similar patterns for future returns. The bad-news D10-D1 strategy produces significantly negative returns for all proxies except for DISP. For the good-news strategy, only the SIZE and COV proxies produce significantly positive returns. Information uncertainty has a slightly greater effect following bad news than following good news.<sup>15</sup> Such asymmetry between good and bad news might be partly explained by short-sale restrictions.

Another interesting observation is that firm size works well as a proxy for information uncertainty. Market participants underreact more to new information for small firms than for large firms. As a result, small firms have relatively lower future returns following bad news and relatively higher future returns following good news. In other words, the size premium (SMB), which is defined as the return differential between five small-size deciles and five big-size deciles in each news category, is positive following good news but negative following bad news. The positive premium following good news offsets the negative premium following bad news, resulting in a positive premium overall. The negative (positive) SMB following bad (good) news is also persistent over time. Following bad news ( $REV < 0$ ), SMB is negative in 17 out of 19 years, with an average annual return of  $-5.56\%$  and a  $t$ -statistic of  $-2.41$  (results untabulated). Following good news, SMB is positive for 15 years, with an average annual return of  $7.41\%$  and a  $t$ -statistic of  $3.09$ . For the whole sample, SMB is  $0.05\%$  and is indistinguishable from zero in the 1983 to 2001 sample period, which is consistent with previous evidence. This evidence might provide an alternative explanation to the well-known size anomaly. Certainly, it is interesting to see whether this approach can fully explain the size anomaly both by examining bad/good news for small firms versus large firms during different market conditions, and by examining whether there is more bad news for small firms in the late 1980s due to competition and globalization.<sup>16</sup> A full investigation of this issue is beyond the scope of this paper.

The opposite effects of information uncertainty on future returns following good and bad news have a big impact on the performance of a trading strategy

<sup>14</sup> In untabulated results, I find that the predictability of stock returns based on information uncertainty lasts for at least 6 months for the bad-news category but only 1 or 2 months for the good-news category. The asymmetry of results between good and bad news might be partially due to short-sale restrictions, especially for high-uncertainty firms. The fact that the predictability of stock returns is much more short-lived for good news than for bad news explains why high-uncertainty firms have slightly lower returns than low-uncertainty firms in the no-news category. The no-news category is actually a combination of good news and bad news from previous months. Given that the information effect lasts longer for bad news, the no-news category exhibits a pattern more similar to the bad-news category than to the good-news category. In this sense, the no-news category is mislabeled.

<sup>15</sup> The asymmetry of results following good versus bad news is relatively big for SIGMA, which is consistent with the results of Ang et al. (2003), who find that idiosyncratic stock volatility is negatively priced in the overall market.

<sup>16</sup> The evidence in Fama and French (1995) that the recession in 1981 and 1982 turns into a prolonged earnings depression for small stocks but not for large stocks supports this argument.

based on analyst forecast revisions. For low-uncertainty firms the initial market reaction is largely complete. For example, a trading strategy of buying good-news stocks and shorting bad-news stocks yields a small 0.46% monthly return when I focus on low-volatility stocks (SIGMA D1, Table III). On the other hand, the initial market response is far from complete for high-uncertainty firms. A similar strategy using high-volatility stocks (SIGMA D10, Table III) generates a 2.39% monthly return. To make a comparison, a simple trading strategy with a short position in all downward-revision stocks and a long position in all upward-revision stocks yields a 1.13% monthly return (Table II, Panel B).

### *C. Portfolio Returns by Price Momentum and Information Uncertainty Proxy*

Intermediate-horizon stock returns offer another measure of the nature of news. Past winners imply good news and past losers imply bad news. Therefore, my prediction is that greater information uncertainty predicts relatively lower future returns for past losers and relatively higher future returns for past winners.

Table IV shows the returns when momentum interacts with information uncertainty. Each month I sort stocks into five quintiles based on past returns from  $t - 11$  to  $t - 1$ .<sup>17</sup> For each momentum quintile, I further sort stocks into five groups by information uncertainty level. As shown in Table IV, information uncertainty is highly negatively correlated with 1-month-ahead stock returns for past losers. For example, the youngest firm (AGE) quintile earns an average  $-0.47\%$  monthly return, compared to  $1.20\%$  for the oldest firm quintile. The return differential between these two quintiles (U5-U1) is  $-1.67\%$  ( $t = -4.42$ ). On the other hand, information uncertainty is strongly positively correlated with 1-month-ahead returns for past winners. The youngest firm quintile gains  $2.43\%$  per month, but the oldest firm quintile gains only  $1.60\%$  per month. The return differential between these two quintiles is  $0.83\%$  ( $t = 2.64$ ). Other information uncertainty proxies produce similar results. This evidence clearly supports my hypothesis.

Table IV also shows that because of a strong interaction effect between momentum and information uncertainty, the momentum effect is much stronger for high-uncertainty firms than for low-uncertainty firms.<sup>18</sup> The return from a trading strategy with a long position in past winners and a short position in past losers increases monotonically as information uncertainty increases. For example, using the AGE proxy, the momentum trading strategy produces an average monthly return as high as  $2.90\%$  ( $t = 7.21$ ) for the youngest firm

<sup>17</sup> I allow a 1-month lag between the momentum measure and the portfolio formation date, consistent with Fama and French (1996) and Diether et al. (2002).

<sup>18</sup> This two-way nonindependent sort by momentum and then by information uncertainty accurately measures the information uncertainty effect within each momentum group but not the momentum effect within each uncertainty group. I replicate the results when first sorting stocks by information uncertainty and then by momentum. In this way, the hedge portfolio returns accurately reflect the momentum profit within each uncertainty group. I also find similar results using independent sorts by momentum and information uncertainty.

Table IV  
Portfolio Returns by Price Momentum and Information  
Uncertainty Proxy

This table reports average monthly portfolio returns sorted by price momentum and information uncertainty proxy. Each month I first sort stocks into five quintiles based on returns from months  $t - 11$  to  $t - 1$ . For each momentum quintile, I further sort stocks into five groups based on information uncertainty proxy. Firm size (MV) is the market capitalization (in millions of dollars) at the end of month  $t$ . Firm age (AGE) is the number of years since the firm was first covered by CRSP. Analyst coverage (COV) is the number of analysts following the firm in the previous year. Forecast dispersion (DISP) is the standard deviation of analyst forecasts in month  $t$  scaled by the prior year-end stock price. Stock volatility (SIGMA) is the standard deviation of weekly market excess returns over the year ending at the end of month  $t$ . Cash flow volatility (CVOL) is the standard deviation of cash flow from operations in the past 5 years (with a minimum of 3 years), where cash flow from operations is earnings before extraordinary items minus total accruals, scaled by average total assets.  $1/MV$ ,  $1/AGE$ , and  $1/COV$  are the reciprocals of MV, AGE, and COV, respectively. Stocks with a price less than \$5 at the portfolio formation date are excluded from the sample. Stocks are held for 1 month, and portfolio returns are equally weighted. The sample period is from January 1983 to December 2001;  $t$ -statistics in parentheses are adjusted for autocorrelation.

	Momentum Quintile					M5 – M1
	M1 (Losers)	M2	M3	M4	M5 (Winners)	
Uncertainty Proxied by 1/MV						
U1 (low)	0.75%	1.09%	1.09%	1.21%	1.53%	0.78% (1.74)
U2	0.50%	1.04%	1.14%	1.36%	1.88%	1.38% (3.53)
U3	0.66%	1.08%	1.23%	1.37%	1.96%	1.30% (3.55)
U4	0.23%	1.15%	1.46%	1.71%	2.33%	2.09% (6.07)
U5 (high)	0.35%	1.11%	1.40%	1.69%	2.58%	2.23% (6.45)
U5 – U1	−0.40% (−1.30)	0.02% (0.08)	0.31% (1.23)	0.48% (1.94)	1.05% (3.96)	1.45% (4.43)
Uncertainty Proxied by 1/AGE						
U1 (low)	1.20%	1.26%	1.19%	1.32%	1.60%	0.40% (1.08)
U2	0.89%	1.21%	1.37%	1.40%	1.90%	1.01% (2.71)
U3	0.52%	1.08%	1.44%	1.57%	2.04%	1.52% (4.03)
U4	0.32%	1.00%	1.24%	1.58%	2.24%	1.92% (5.08)
U5 (high)	−0.47%	0.86%	1.05%	1.47%	2.43%	2.90% (7.21)
U5 – U1	−1.67% (−4.42)	−0.41% (−1.55)	−0.14% (−0.53)	0.15% (0.54)	0.83% (2.64)	2.50% (7.98)
Uncertainty Proxied by 1/COV						
U1 (low)	0.72%	1.08%	1.09%	1.17%	1.64%	0.92% (2.13)
U2	0.74%	0.98%	1.19%	1.28%	1.88%	1.13% (2.85)

(continued)

Table IV—Continued

	Momentum Quintile					M5 – M1
	M1 (Losers)	M2	M3	M4	M5 (Winners)	
U3	0.31%	1.05%	1.15%	1.46%	1.93%	1.63% (4.69)
U4	0.21%	0.75%	1.13%	1.29%	2.24%	2.03% (5.31)
U5 (high)	−0.12%	0.96%	1.14%	1.34%	2.02%	2.14% (5.85)
U5 – U1	−0.84% (−3.00)	−0.12% (−0.57)	0.05% (0.25)	0.17% (0.50)	0.38% (1.44)	1.22% (4.17)
Uncertainty Proxied by DISP						
U1 (low)	0.65%	1.00%	1.22%	1.42%	1.76%	1.11% (2.62)
U2	0.95%	1.08%	1.18%	1.35%	2.00%	1.05% (2.90)
U3	0.66%	1.29%	1.30%	1.50%	1.92%	1.26% (3.31)
U4	0.39%	1.12%	1.26%	1.46%	2.09%	1.70% (4.51)
U5 (high)	0.15%	1.03%	1.41%	1.63%	2.45%	2.30% (6.76)
U5 – U1	−0.50% (−1.80)	0.02% (0.13)	0.18% (0.93)	0.22% (1.15)	0.69% (3.20)	1.19% (4.02)
Uncertainty Proxied by SIGMA						
U1 (low)	1.11%	1.37%	1.36%	1.51%	1.75%	0.63% (2.04)
U2	0.98%	1.32%	1.34%	1.50%	1.97%	1.00% (2.85)
U3	0.61%	1.09%	1.39%	1.47%	2.06%	1.45% (3.65)
U4	0.12%	1.03%	1.25%	1.46%	2.23%	2.10% (4.84)
U5 (high)	−0.35%	0.65%	0.95%	1.30%	2.28%	2.63% (5.91)
U5 – U1	−1.47% (−3.04)	−0.72% (−1.84)	−0.41% (−0.98)	−0.21% (−0.47)	0.53% (1.01)	2.00% (5.62)
Uncertainty Proxied by CVOL						
U1 (low)	1.04%	1.28%	1.33%	1.27%	1.72%	0.68% (1.85)
U2	0.98%	1.16%	1.27%	1.39%	1.72%	0.74% (2.00)
U3	0.81%	0.92%	1.17%	1.47%	2.05%	1.24% (3.44)
U4	0.56%	1.06%	1.16%	1.31%	2.22%	1.67% (3.97)
U5 (high)	0.06%	0.56%	1.19%	1.38%	2.11%	2.05% (5.18)
U5 – U1	−0.97% (−2.92)	−0.73% (−2.45)	−0.14% (−0.41)	0.11% (0.34)	0.40% (1.13)	1.37% (4.74)

quintile, compared to only 0.40% ( $t = 1.08$ ) for the oldest firm quintile. The return differential between these two momentum strategies is 2.50% ( $t = 7.98$ ). The return differential between the momentum strategies for high-uncertainty stocks (U5) and low-uncertainty stocks (U1) ranges from 1.19% to 2.00% per month for other information uncertainty proxies and is highly significant in each case.

The above results are based on five price momentum portfolios and five uncertainty portfolios ( $5 \times 5$ ). My results are not specific to this partitioning. In analyses untabulated, I use 10 price momentum and 3 uncertainty portfolios ( $10 \times 3$ ) and use 3 price momentum and 10 uncertainty portfolios ( $3 \times 10$ ). Generally, the uncertainty effect is as strong as that reported in Table IV. For each proxy, the return differential between the momentum strategy for high-uncertainty stocks and that for low-uncertainty stocks is positive and significantly different from zero. These results are consistent with the underreaction explanation for the momentum phenomenon, in the sense that investors underreact to a higher degree when there is greater information uncertainty.

#### *D. Portfolio Returns by Forecast Revision, Momentum, and Information Uncertainty Proxy*

The final portfolio strategy uses a four-way sort by forecast revision, momentum, and two information uncertainty proxies. Double sorts by analyst forecast revision and momentum should better identify firms with really bad and really good news, and therefore provide a more precise test of the effect of information uncertainty on investor underreaction behavior. I focus on the really bad-news groups (losers with downward revisions) and really good-news groups (winners with upward revisions) in the test. On the information uncertainty side, I sort by size and each of the other information uncertainty proxies because firm size is extensively studied in the prior literature.

To form portfolios, I first sort stocks into three categories based on forecast revisions in the current month. Within each revision category, I sort the stocks into three groups based on past returns from  $t - 11$  to  $t - 1$ . Then for each revision and momentum group, I further sort the stocks into three divisions by size, and finally into three uncertainty subsets. This four-way sort classifies stocks into 81 portfolios. For each month there is an average of 24, 38, and 19 stocks in each portfolio for the negative, zero, and positive revision categories, respectively.

Table V presents the 1-month-ahead returns for the really bad- and really good-news groups. Following bad news, stock returns monotonically decrease in each size group as information uncertainty increases for most categories. The return differential between high- and low-uncertainty firms (U3-U1) is significantly negative for 8 out of 15 size groups. After good news, high-uncertainty firms have higher future returns than do low-uncertainty ones. The U3-U1 strategy yields positive returns in all but one size group. These results provide further support for my hypothesis. Consistent with the evidence in Table III, the size premiums are uniformly negative following bad news and highly posi-

Table V  
Portfolio Returns by Forecast Revision, Momentum, and Information Uncertainty Proxy

This table reports average monthly returns for really bad (losers with  $REV < 0$ ) and really good (winners with  $REV > 0$ ) portfolios, using double sorts by the nature of news and double sorts by information uncertainty proxies. Each month I sort stocks into three categories depending on whether the forecast revision is negative, zero, or positive. The forecast revision is the average of individual forecast revisions by analysts who covered the firm in both month  $t - 1$  and  $t$ . For each category, I sort stocks into three groups based on returns from months  $t - 11$  to  $t - 1$ . For each news group, I further sort stocks into three divisions by firm size, and finally into three subsets based on another uncertainty proxy. Firm size (MV) is the market capitalization (in millions of dollars) at the end of month  $t$ . Firm age (AGE) is the number of years since the firm was first covered by CRSP. Analyst coverage (COV) is the number of analysts following the firm in the previous year. Forecast dispersion (DISP) is the standard deviation of analyst forecasts in month  $t$  scaled by the prior year-end stock price. Stock volatility (SIGMA) is the standard deviation of weekly market excess returns over the year ending at the end of month  $t$ . Cash flow volatility (CVOL) is the standard deviation of cash flow from operations in the past 5 years (with a minimum of 3 years), where cash flow from operations is earnings before extraordinary items minus total accruals, scaled by average total assets.  $1/MV$ ,  $1/AGE$ , and  $1/COV$  are the reciprocals of MV, AGE, and COV, respectively. Stocks with a price less than \$5 at the portfolio formation date are excluded from the sample. Stocks are held for 1 month, and portfolio returns are equally weighted. The sample period is from January 1983 to December 2001;  $t$ -statistics in parentheses are adjusted for autocorrelation.

	Bad News (Losers with $REV < 0$ )				Good News (Winners with $REV > 0$ )			
	Small Cap		Mid Cap		Small Cap		Mid Cap	
	Large Cap	Small-Large	Large Cap	Small-Large	Large Cap	Small-Large	Large Cap	Small-Large
Uncertainty Proxied by 1/AGE								
U1	0.37%	1.21%	1.10%	-0.73% (-2.14)	2.69%	1.92%	1.72%	0.97% (3.27)
U2	0.16%	0.64%	1.18%	-1.02% (-2.64)	3.39%	2.25%	1.81%	1.58% (4.07)
U3	-0.93%	-0.24%	0.50%	-1.43% (-3.77)	3.57%	2.86%	2.42%	1.15% (3.03)
U3 - U1	-1.30% (-3.15)	-1.45% (-3.59)	-0.60% (-1.64)		0.88% (3.26)	0.94% (2.09)	0.70% (2.01)	
Uncertainty Proxied by 1/COV								
U1	0.09%	0.72%	0.71%	-0.62% (-1.58)	3.24%	2.08%	2.00%	1.24% (3.25)
U2	0.16%	0.54%	1.04%	-0.88% (-2.42)	3.03%	2.65%	2.13%	0.90% (2.52)
U3	-0.79%	0.15%	0.78%	-1.57% (-4.86)	3.71%	2.09%	1.93%	1.78% (4.93)
U3 - U1	-0.88% (-2.68)	-0.57% (-1.51)	0.07% (0.23)		0.47% (1.42)	0.00% (0.13)	-0.06% (-0.06)	
Good News - Bad News								
U1					2.32% (5.34)	0.71% (1.58)	3.15% (7.06)	1.36% (2.85)
U2					3.23% (7.52)	1.61% (3.19)	2.87% (6.81)	2.11% (4.51)
U3					4.50% (10.32)	3.10% (6.30)	4.50% (10.37)	1.94% (4.17)
U3 - U1								

Information Uncertainty and Stock Returns

Uncertainty Proxied by DISP											
U1	0.55%	0.79%	0.84%	-0.29% (-0.73)	2.89%	2.19%	1.71%	1.18% (3.46)	2.34% (5.23)	1.40% (2.79)	0.87% (1.83)
U2	-0.18%	0.69%	0.84%	-1.02% (-2.88)	3.59%	2.06%	2.21%	1.38% (3.80)	3.77% (8.86)	1.37% (3.33)	1.37% (3.15)
U3	-0.47%	0.41%	0.98%	-1.45% (-4.00)	3.13%	2.43%	1.98%	1.15% (3.13)	3.60% (8.44)	2.02% (4.43)	1.00% (2.64)
U3 - U1	-1.02% (-2.89)	-0.38% (-1.14)	0.15% (0.54)		0.24% (1.05)	0.24% (0.83)	0.27% (1.05)				
Uncertainty Proxied by SIGMA											
U1	0.51%	0.99%	1.32%	-0.81% (-2.76)	2.75%	2.07%	1.72%	1.03% (3.64)	2.24% (6.45)	1.08% (3.07)	0.40% (1.32)
U2	0.01%	0.66%	0.98%	-0.97% (-2.68)	3.20%	2.16%	1.86%	1.34% (3.92)	3.19% (7.26)	1.50% (3.10)	0.88% (1.93)
U3	-0.90%	-0.01%	0.47%	-1.37% (-3.29)	3.81%	2.66%	2.35%	1.46% (3.77)	4.71% (9.70)	2.67% (4.56)	1.88% (3.80)
U3 - U1	-1.41% (-3.00)	-1.00% (-2.04)	-0.85% (-2.22)		1.06% (2.03)	0.59% (1.13)	0.63% (1.57)				
Uncertainty Proxied by CVOL											
U1	0.50%	0.92%	1.13%	-0.62% (-1.76)	3.00%	1.81%	1.73%	1.27% (3.23)	2.53% (5.97)	0.87% (2.16)	0.58% (1.41)
U2	0.37%	1.16%	0.90%	-0.53% (-1.48)	3.01%	1.93%	1.91%	1.10% (2.81)	2.64% (5.39)	0.77% (1.73)	1.01% (2.26)
U3	-0.25%	0.46%	0.56%	-0.81% (-2.06)	3.82%	2.42%	2.50%	1.32% (3.39)	4.07% (8.25)	1.96% (3.50)	1.95% (3.94)
U3 - U1	-0.74% (-2.01)	-0.46% (-1.36)	-0.56% (-1.87)		0.82% (1.86)	0.61% (1.49)	0.77% (2.31)				



tive following good news, and they are significantly different from zero in most uncertainty groups.

Furthermore, firm size and other information uncertainty proxies interact in a plausible way. The uncertainty effect is greatest for the smallest size group in both good- and bad-news categories, indicating that other information uncertainty proxies play a more significant role for smaller firms. Similarly, the size effect is typically the strongest for the highest uncertainty group following good or bad news. In untabulated results, I find that uncertainty proxies for the bad-news category closely match those for the good-news category in each case, which implies that information uncertainty alone cannot explain the observed return pattern.

The double sorts in each dimension do not subsume each other, which suggests that each sort has incremental information and no proxy is perfect. A trading strategy that uses this categorization achieves remarkable returns. For example, for high-uncertainty stocks (small size and young age), the trading strategy of buying past winners with upward revisions and shorting past losers with downward revisions generates an average 4.50% ( $t = 10.32$ ) monthly return. The same strategy yields 2.32% ( $t = 5.34$ ), 1.92% ( $t = 4.09$ ), and 0.62% ( $t = 1.66$ ) for small firms with a long history, large firms with a short history, and large firms with a long history, respectively. These results indicate that market reaction to new information is quite complete for low-uncertainty firms but not for high-uncertainty firms, and that size and other proxies for information uncertainty have similar effects on investor underreaction but do not subsume each other.

#### **IV. Four-Factor Model Results**

In this section, I examine whether my information uncertainty results can be explained using a rational approach. Fama and French (1996) show that their three-factor model ( $Rm - Rf$ , SMB, and HML) can explain most commonly documented Capital Asset Pricing Model (CAPM) anomalies except for the continuation of short-term returns. They argue that the three-factor model works like an equilibrium pricing model in the spirit of Merton's (1973) intertemporal CAPM or Ross's (1976) arbitrage pricing theory and that SMB and HML mimic combinations of two underlying risk factors or state variables of special hedging concern to investors. Empirically, SMB represents the size premium and equals the return differential between portfolios of small and large stocks. Similarly, HML represents the value premium and equals the return differential between portfolios of stocks with high book-to-market ratios and low book-to-market ratios (see Fama and French (1996) for details on these three factors).

Since the Fama–French three-factor model does not capture the momentum effect, I use a four-factor model (e.g., Carhart (1997)) to test portfolio returns. If the four-factor model can capture the cross-sectional variation in stock returns, then the intercept from the following regression should be statistically indistinguishable from zero,

$$R_{it} - R_{ft} = \alpha + b_{iM}(R_{Mt} - R_{ft}) + s_i \text{SMB}_t + h_i \text{HML}_t + m_i \text{UMD}_t + \varepsilon_{it}, \quad (1)$$

where  $R_{it} - R_{ft}$  is the return of portfolio  $i$  in excess of the risk-free rate in month  $t$ ,  $R_{Mt} - R_{ft}$  is the excess return of the market value-weighted portfolio, and UMD is the return difference between portfolios of past winners and past losers.<sup>19</sup>

Table VI reports the intercepts of the four-factor model for 15 portfolios for each uncertainty proxy. Each month I sort stocks into three categories depending on whether the forecast revision in the past month is negative, zero, or positive. For each category, I further sort stocks into five portfolios based on an information uncertainty proxy. The intercepts from the four-factor model are uniformly negative for bad-news portfolios and positive for good-news portfolios. More importantly, the magnitude of the intercept is positively related to the level of uncertainty, which implies that high-uncertainty portfolios earn more negative abnormal returns following bad news and more positive abnormal returns following good news in a four-factor world. For example, young firm portfolios have intercepts of  $-0.709$  ( $t = -4.18$ ) following bad news and  $1.176$  ( $t = 7.29$ ) following good news, which correspond to  $-8.51\%$  and  $14.11\%$  annual abnormal returns, respectively. A trading strategy with a short position in young firms with downward revisions and a long position in young firms with upward revisions generates a  $22.62\%$  annual abnormal return after controlling for the market, size, value, and momentum effects. For no-news portfolios, the intercepts are indistinguishable from zero in most cases. This pattern of intercepts from the four-factor model further confirms my hypothesis.

Untabulated results show that the risk loadings on  $R_{Mt} - R_{ft}$ , SMB, HML, and UMD are as expected. The risk loadings on the market premium are each close to one for all 90 portfolios with  $t$ -statistics over 30. High-uncertainty portfolios have higher loadings on SMB, suggesting that high-uncertainty firms tend to be small. The loadings on HML are typically lower for high-uncertainty stocks except when information uncertainty is proxied by firm size or analyst forecast dispersion, which suggests that high-uncertainty firms are more likely to be growth firms. The risk loadings on UMD are uniformly negative for bad-news portfolios but usually positive for good-news portfolios, confirming momentum as a proxy for the nature of news. The adjusted  $R^2$  is around 0.9 across portfolios, suggesting that the four-factor model has reasonable explanatory power.

In summary, the level of uncertainty is positively (negatively) related to abnormal stock returns following good (bad) news. Although each proxy might also capture other risk factors or contain substantial measurement error, consistent results across different proxies lend strong support to the view that information uncertainty magnifies behavioral biases and is not a priced risk factor. Because the information uncertainty proxies consistently increase returns following good news but decrease returns following bad news, it is difficult to construct a risk-based story for this effect. Finally, the inclusion of a

<sup>19</sup> I construct UMD in the same way as in Carhart (1997), and download the Fama-French three factors from Ken French's website: [http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data\\_library.html](http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html).

Table VI  
Four-Factor Model Results

This table reports the intercepts of the four-factor regression model for monthly excess returns of the information uncertainty quintiles for three news categories based on analyst forecast revisions. The model estimated is

$$R_{it} - R_{ft} = a + b_{iM}(R_{Mt} - R_{ft}) + s_i \text{SMB}_t + h_i \text{HML}_t + m_i \text{UMD}_t + \varepsilon_{it},$$

where  $R_{Mt} - R_{ft}$ , SMB, and HML are as defined in Fama and French (1996), and UMD is momentum as defined in Carhart (1997). Each month I sort stocks into three categories depending on whether the forecast revision is negative (bad news), zero (no news), or positive (good news) in month  $t - 1$ . For each news category, I further sort stocks into five portfolios based on an information uncertainty proxy. Firm size (MV) is the market capitalization (in millions of dollars) at the end of month  $t - 1$ . Firm age (AGE) is the number of years since the firm was first covered by CRSP. Analyst coverage (COV) is the number of analysts following the firm in the previous year. Forecast dispersion (DISP) is the standard deviation of analyst forecasts in month  $t$  scaled by the prior year-end stock price. Stock volatility (SIGMA) is the standard deviation of weekly market excess returns over the year ending at the end of month  $t - 1$ . Cash flow volatility (CVOL) is the standard deviation of cash flow from operations in the past 5 years (with a minimum of 3 years), where cash flow from operations is earnings before extraordinary items minus total accruals, scaled by average total assets. 1/MV, 1/AGE, and 1/COV are the reciprocals of MV, AGE, and COV, respectively. The sample period is from January 1983 to December 2001. Stocks with a price less than \$5 at the portfolio formation date are excluded from the sample, and White heteroskedasticity-adjusted  $t$ -statistics are in parentheses.

News	Uncertainty	1/MV	1/AGE	1/COV	DISP	SIGMA	CVOL
Bad News (REV < 0)	Q1 (Low)	-0.173	-0.152	-0.012	-0.185	-0.066	-0.153
		(-1.80)	(-1.10)	(-0.69)	(-0.89)	(-0.47)	(-1.00)
	Q2	-0.092	-0.102	-0.123	-0.074	-0.319	-0.088
		(-0.48)	(-0.64)	(-0.77)	(-0.36)	(-2.29)	(-0.55)
	Q3	-2.85	-2.212	-0.285	-0.362	-0.457	-0.244
		(-1.60)	(-1.33)	(-1.82)	(-2.28)	(-2.95)	(-3.33)
	Q4	-0.370	-0.318	-0.319	-0.244	-0.240	-0.214
		(-2.26)	(-1.82)	(-1.99)	(-1.75)	(-1.42)	(-1.16)
	Q5 (high)	-0.519	-0.709	-0.724	-0.415	-0.361	-4.410
		(-3.17)	(-4.18)	(-5.62)	(-2.50)	(-1.57)	(-2.10)
No News (REV = 0)	Q1 (Low)	-0.210	-0.139	0.161	-0.003	0.157	0.082
		(-1.86)	(-0.96)	(1.40)	(-0.51)	(1.17)	(0.64)
	Q2	0.008	0.317	0.226	0.206	0.032	0.082
		(0.10)	(2.33)	(1.76)	(1.76)	(0.20)	(0.60)
	Q3	0.186	0.292	0.077	0.241	0.138	0.357
		(1.84)	(2.89)	(0.72)	(1.69)	(1.10)	(2.67)
	Q4	0.451	0.148	0.183	0.321	0.456	0.263
		(2.94)	(1.09)	(1.58)	(2.89)	(3.44)	(1.84)
	Q5 (high)	0.483	0.270	0.156	0.137	0.116	0.219
		(2.81)	(1.62)	(1.10)	(1.62)	(0.59)	(1.10)
Good News (REV > 0)	Q1 (Low)	0.169	0.074	0.419	0.369	0.210	0.250
		(1.91)	(0.57)	(3.76)	(3.05)	(1.82)	(1.53)
	Q2	0.236	0.251	0.170	0.767	0.198	0.329
		(1.65)	(1.86)	(1.15)	(4.78)	(1.74)	(2.60)
	Q3	0.284	0.675	0.685	0.664	0.363	0.624
		(1.66)	(4.47)	(4.60)	(4.76)	(2.28)	(3.75)
	Q4	0.888	0.892	0.751	0.587	0.791	0.766
		(5.57)	(5.72)	(4.85)	(4.02)	(4.71)	(3.87)
	Q5 (high)	1.591	1.176	1.087	0.588	1.596	1.318
		(9.20)	(7.29)	(7.52)	(3.40)	(7.24)	(6.81)

size factor in the four-factor model cannot subsume the abnormal stock returns based on size and the nature of news, confirming early evidence that firm size is more likely to be associated with information uncertainty rather than being a common risk factor in the cross-section of stock returns.

## **V. Market Reaction to Subsequent Earnings Announcements**

The evidence in the previous sections indicates that post-news price drift increases with information uncertainty. A limitation of my previous analyses is that this relationship could be attributable to unidentified risk factors or unknown research design flaws. This section mitigates these concerns by examining stock price reactions to earnings announcements after the portfolio formation date. Because daily expected returns are close to zero, the model used for expected returns does not have a large effect on inferences about abnormal returns (Fama (1998)). Therefore, risk-based models would predict zero returns over this short window. If investor behavior exhibits underreaction to news related to future earnings, investors should correct their misvaluations around subsequent earnings announcement dates. Therefore, we should observe a positive relation between the nature of news and the stock price reactions to the subsequent earnings announcement (Chan, Jegadeesh, and Lakonishok (1999)). In particular, we expect to see a negative market reaction on the earnings announcement date following bad news and a positive one following good news. If information uncertainty exacerbates an investor's behavioral bias, we expect to see more positive (negative) reactions following good (bad) news for high-uncertainty stocks than for low-uncertainty stocks.

Since earnings are announced on a quarterly basis, I form five uncertainty portfolios for each calendar quarter following bad and good news, respectively. Good news refers to upward analyst forecast revisions in the previous month or past winners (top quintile) and vice versa for bad news. Following Bernard and Thomas (1990) and Jegadeesh and Titman (1993), the announcement period for each quarterly announcement is defined as the 3-day period beginning two days prior to the Compustat earnings announcement date.

Table VII presents the average daily market excess returns (measured as raw return minus the contemporaneous value-weighted market return) from the announcement period tests. Panel A reports the results when the nature of news is based on analyst forecast revisions, and in Panel B it is based on the past 11-month stock returns. Both panels show that the 3-day excess returns around earnings announcements are predictable. The signs and magnitudes of the excess returns are consistent with my hypothesis. The market reaction to earnings announcements is negative for bad-news portfolios and positive for good-news portfolios for all uncertainty proxies. More importantly, the magnitude of excess returns around the quarterly earnings announcement date increases with the level of information uncertainty. In both panels, a zero-investment portfolio with a long position in good-news stocks and a short position in bad-news stocks generates the highest returns for high-uncertainty stocks for both types of news measures and for all proxies except for the AGE/REV combination.

Table VII  
Excess Returns Around a 3-Day Earnings Announcement Window

This table reports average daily excess returns around a 3-day earnings announcement window. Excess returns are measured as raw returns minus the value-weighted market return. The 3-day window starts two days prior to the Compustat earnings announcement date. Each quarter I sort stocks into bad and good news categories. For each category, I further sort stocks into five portfolios based on information uncertainty. In Panel A, bad news refers to downward analyst forecast revisions (REV), and good news refers to upward analyst forecast revisions in month  $t$ . In Panel B, bad news refers to the bottom momentum quintile (past losers), and good news refers to the top momentum quintile (past winners), where momentum is the accumulated return from month  $t - 11$  to  $t - 1$ . GMB is a zero-investment portfolio with a long position in good-news stocks and a short position in bad-news stocks. Firm size (MV) is the market capitalization (in millions of dollars) at the end of month  $t$ . Firm age (AGE) is the number of years since the firm was first covered by CRSP. Analyst coverage (COV) is the number of analysts following the firm in the previous year. Forecast dispersion (DISP) is the standard deviation of analyst forecasts in month  $t$  scaled by the prior year-end stock price. Stock volatility (SIGMA) is the standard deviation of weekly market excess returns over the year ending at the end of month  $t$ . Cash flow volatility (CVOL) is the standard deviation of cash flow from operations in the past 5 years (with a minimum of 3 years), where cash flow from operations is earnings before extraordinary items minus total accruals, scaled by average total assets. 1/MV, 1/AGE, and 1/COV are the reciprocals of MV, AGE, and COV, respectively. Stocks with a price less than 5 dollars at the portfolio formation date are excluded from the sample. Portfolio returns are equally weighted. The sample period is from January 1983 to December 2001.

Panel A. Bad News—Downward Revisions (REV < 0), Good News—Upward Revisions (REV > 0)									
	Sorted by 1/MV			Sorted by 1/AGE			Sorted by 1/COV		
	REV < 0	REV > 0	GMB	REV < 0	REV > 0	GMB	REV < 0	REV > 0	GMB
Q1 (low)	0.02%	0.13%	0.11% (3.27)	−0.01%	0.09%	0.10% (3.95)	0.05%	0.14%	0.09% (1.97)
Q2	−0.03%	0.16%	0.18% (4.39)	−0.01%	0.17%	0.18% (4.91)	−0.03%	0.11%	0.14% (3.01)
Q3	−0.02%	0.21%	0.23% (4.42)	−0.03%	0.25%	0.27% (5.14)	−0.13%	0.23%	0.35% (6.81)
Q4	−0.08%	0.24%	0.31% (6.37)	−0.12%	0.31%	0.43% (8.08)	−0.07%	0.26%	0.33% (5.18)
Q5 (high)	−0.11%	0.35%	0.46% (8.34)	−0.07%	0.28%	0.35% (7.50)	−0.11%	0.28%	0.38% (7.72)
Q5 − Q1	−0.13% (−3.23)	0.22% (4.33)		−0.06% (−1.15)	0.19% (3.10)		−0.16% (−3.80)	0.14% (2.63)	

	Sorted by DISP			Sorted by SIGMA			Sorted by CVOL		
	REV < 0	REV > 0	GMB	REV < 0	REV > 0	GMB	REV < 0	REV > 0	GMB
Q1 (low)	−0.07%	0.20%	0.27% (4.56)	0.01%	0.12%	0.11% (3.62)	−0.06%	0.14%	0.19% (5.48)
Q2	−0.01%	0.19%	0.19% (4.60)	−0.01%	0.14%	0.15% (4.07)	−0.04%	0.15%	0.18% (4.34)
Q3	0.00%	0.22%	0.22% (5.38)	−0.08%	0.26%	0.34% (5.10)	−0.03%	0.24%	0.27% (4.72)
Q4	−0.03%	0.24%	0.28% (5.48)	−0.06%	0.23%	0.29% (4.92)	−0.05%	0.22%	0.27% (5.02)
Q5 (high)	−0.10%	0.24%	0.33% (5.80)	−0.08%	0.32%	0.40% (5.62)	−0.11%	0.36%	0.47% (6.24)
Q5 − Q1	−0.02% (−0.42)	0.04% (0.76)		−0.09% (−1.40)	0.20% (2.75)		−0.05% (−0.93)	0.22% (2.95)	

(continued)

A comparison between Table IV and Table VII, Panel B reveals additional evidence. Table IV reports average monthly returns and Table VII, Panel B reports average daily returns. Given that a month typically has at least 20 trading days, the announcement period reactions represent a disproportionate share of

Table VII—Continued

Panel B. Bad News—Past Losers, Good News—Past Winners									
	Sorted by 1/MV			Sorted by 1/AGE			Sorted by 1/COV		
	Losers	Winners	GMB	Losers	Winners	GMB	Losers	Winners	GMB
Q1 (low)	0.00%	0.19%	0.19% (2.97)	−0.04%	0.16%	0.21% (3.94)	0.00%	0.17%	0.18% (2.35)
Q2	−0.02%	0.19%	0.21% (3.02)	−0.06%	0.21%	0.26% (4.01)	−0.09%	0.22%	0.31% (5.17)
Q3	−0.14%	0.18%	0.31% (6.27)	−0.14%	0.20%	0.34% (5.22)	−0.03%	0.19%	0.22% (3.31)
Q4	−0.07%	0.26%	0.33% (4.98)	−0.03%	0.24%	0.27% (4.07)	−0.12%	0.20%	0.33% (3.89)
Q5 (high)	−0.11%	0.34%	0.44% (6.82)	−0.07%	0.31%	0.38% (5.03)	−0.10%	0.30%	0.40% (6.41)
Q5 − Q1	−0.11% (−1.90)	0.15% (2.19)		−0.03% (−0.52)	0.14% (2.03)		−0.10% (−1.67)	0.12% (1.87)	

	Sorted by DISP			Sorted by SIGMA			Sorted by CVOL		
	Losers	Winners	GMB	Losers	Winners	GMB	Losers	Winners	GMB
Q1 (low)	0.00%	0.17%	0.17% (2.46)	−0.04%	0.17%	0.21% (4.23)	−0.01%	0.17%	0.18% (2.72)
Q2	0.04%	0.21%	0.17% (2.43)	−0.04%	0.23%	0.27% (4.24)	−0.09%	0.22%	0.31% (4.39)
Q3	−0.11%	0.23%	0.34% (3.76)	−0.05%	0.25%	0.30% (4.36)	−0.05%	0.27%	0.32% (4.60)
Q4	−0.08%	0.21%	0.29% (4.73)	−0.07%	0.26%	0.33% (4.83)	−0.04%	0.27%	0.32% (3.93)
Q5 (high)	−0.14%	0.26%	0.40% (6.81)	−0.13%	0.26%	0.39% (5.40)	−0.14%	0.24%	0.38% (5.42)
Q5 − Q1	−0.14% (−1.94)	0.09% (1.76)		−0.09% (−1.16)	0.09% (1.48)		−0.13% (−2.03)	0.06% (1.09)	

the drift. For example, a zero-investment portfolio on the top SIGMA quintile has a 2.63% monthly return in Table IV. The average of 0.39% daily returns in Table VII, Panel B means that on average at least 15%  $[(0.39 \times 3)/(2.63 \times 3)]$  of the predictable stock returns are concentrated around subsequent earnings announcement dates, which accounts for only 5%  $[=3/(20 \times 3)]$  of trading days.

Overall, the signs and relative magnitudes of the excess returns around subsequent earnings announcement dates are in general accordance with my hypothesis. Given that expected returns should be trivial on a daily basis, this analysis presents more direct evidence that short-term price continuation anomalies are rooted in a failure of information to flow completely into stock prices rather than being driven by missing risk factors.

VI. Robustness Checks

A. Characteristics of Various Trading Strategies

The previous sections show that trading strategies based on the nature of news and/or the level of uncertainty yield significant returns. Table VIII provides a summary of returns from various trading strategies. The Fama–French factors ( $R_m - R_f$ , SMB, and HML) are defined in Section IV and in Fama and

Table VIII  
Characteristics of a Variety of Trading Portfolios

This table presents characteristics of various trading strategies discussed in Section III. The Fama–French factors are as defined in Fama and French (1996). The momentum strategy (MOM) has a short position in past losers and a long position in past winners and is defined in Table IV. The trading strategy (REV\_MOM) has a short position in past losers with downward revisions and a long position in past winners with upward revisions and is defined in Table V. Portfolios with less than five stocks in either the short or long position are deleted. M is the number of months with available portfolio returns from 1983 to 2001. M\_neg is the number of months with negative returns. RET is average monthly portfolio return over the sample period. Sharpe is the Sharpe ratio, defined as the mean return divided by its standard deviation. VOL\_long and VOL\_short are the average monthly dollar trading volume (in millions of dollars) for each stock, and N\_long and N\_short are the average number of stocks in the long and short positions, respectively. Firm size (MV) is the market capitalization (in millions of dollars) at the end of month  $t$ . Firm age (AGE) is the number of years since the firm was first covered by CRSP. Analyst coverage (COV) is the number of analysts following the firm in the previous year. Forecast dispersion (DISP) is the standard deviation of analyst forecasts in month  $t$  scaled by the prior year-end stock price. Stock volatility (SIGMA) is the standard deviation of weekly market excess returns over the year ending at the end of month  $t$ . Cash flow volatility (CVOL) is the standard deviation of cash flow from operations in the past 5 years (with a minimum of 3 years), where cash flow from operations is earnings before extraordinary items minus total accruals, scaled by average total assets.

Trading Strategy	M	M_neg	RET	Sharpe	Fama—French Three Factors			
					VOLM_long	VOLM_short	N_long	N_short
Rm-Rf	228	87	0.71%	0.16				
SMB	228	121	−0.06%	−0.02				
HML	228	101	0.38%	0.11				
The Momentum Strategy (MOM) by Shorting Past Losers and Buying Past Winners (Table IV)								
All stocks	228	68	1.55%	0.28	352	135	422	422
The bottom MV quintile	228	60	2.23%	0.43	13	5	84	84
The bottom AGE quintile	228	64	2.90%	0.46	288	59	84	85
The bottom COV quintile	222	60	2.18%	0.38	40	12	82	80
The top DISP quintile	228	66	2.29%	0.42	221	80	72	72
The top SIGMA quintile	228	72	2.63%	0.38	241	89	84	84
The top CVOL quintile	228	69	2.05%	0.33	375	131	61	61
The Trading Strategy (REV_MOM) by Shorting Past Losers with Downward Revisions and Buying Past Winners with Upward Revisions (Table V)								
Small AGE and small cap	217	53	4.51%	0.68	42	19	19	25
Small COV and small cap	213	42	4.50%	0.68	28	10	18	24
Big DISP and small cap	221	51	3.68%	0.55	41	20	18	23
Big SIGMA and small cap	221	52	4.73%	0.62	52	25	19	24
Big CVOL and small cap	220	63	4.05%	0.53	48	22	14	19



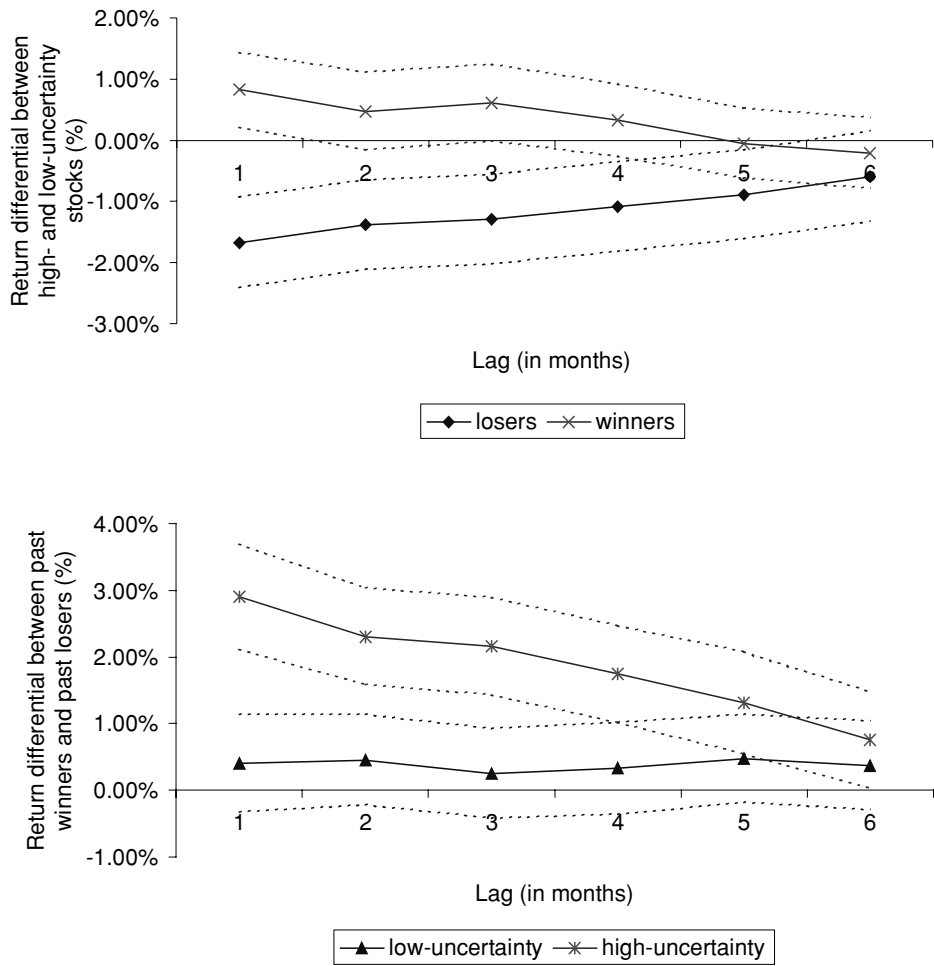
French (1996). The momentum trading strategy (MOM) has a short position in past losers and a long position in past winners and is defined in Table IV. The trading strategy (REV\_MOM) has a short position in past losers with downward revisions and a long position in past winners with upward revisions as defined in Table V. Table VIII shows the results from MOM and REV\_MOM strategies using high-uncertainty stocks. The average market excess return is 0.71% per month with a Sharpe ratio of 0.16. Both the average return and the Sharpe ratio are smaller for HML and are slightly negative for SMB. The MOM strategy for the whole sample yields a 1.55% monthly return with a Sharpe ratio of 0.28. When trading is limited to high-uncertainty stocks, the strategy produces significantly higher returns and Sharpe ratios. The return ranges from 2.05% to 2.90% and the Sharpe ratio ranges from 0.33 to 0.46 for six information uncertainty proxies. Both returns and Sharpe ratios are even higher when the REV\_MOM strategy is used.<sup>20</sup> Additionally, REV\_MOM yields negative returns (N\_neg) in fewer months than the MOM strategy.

In short, the trading strategies based on the nature of news and the level of uncertainty produce impressive average returns and Sharpe ratios, supporting my hypothesis. Certainly, I cannot definitely rule out the possibility that some risk-based model might explain the returns of these strategies, but given the high Sharpe ratios of the portfolios, such a model, based on the arguments of Hansen and Jagannathan (1991), would require investors to have very peculiar preferences.

### *B. Lag in Portfolio Formation*

To examine the persistence of the information uncertainty effect, I duplicate the analysis but wait several months before assigning stocks to portfolios to see how long it takes the market to react completely to the news. Figure 1 shows the effect of uncertainty following good and bad news and the momentum effect among high- and low-uncertainty stocks when I use firm age as the information uncertainty proxy (other proxies produce similar patterns). As the lag increases and uncertainty is resolved, the magnitude of return differentials between high and low uncertainty stocks decreases. The return differential disappears after 6 months for bad news and 1 month for good news. As previously discussed, the persistence of negative returns in the bad-news case is probably due to short-sale restrictions. The fact that the uncertainty effect is

<sup>20</sup> Two caveats about the REV\_MOM strategy are in order. First, as we push for a higher return, the size of the zero-investment portfolio becomes an issue. Certainly, a monthly return of 4.73% and a Sharpe ratio of 0.62 are not achievable for a multibillion-dollar fund, although the investment opportunity is still attractive. Assuming a position in any stock is 5% of average monthly trading volume, the portfolio size of the REV\_MOM strategy on the SIGMA/SIZE combination would be \$60 million. The reason is as follows. The short position has an upper bound of \$30 million ( $=24 \times 25 \times 5\%$ ), while the long position has an upper bound of \$49.4 million ( $=19 \times 52 \times 5\%$ ). The short position is constrained in a zero-investment portfolio. Second, investors may not be able to diversify idiosyncratic risks. Because I sort stocks into 81 portfolios and the REV\_MOM strategy focuses on two portfolios only, the short or long position has only about 20 stocks.



**Figure 1. Lag in portfolio formation.** At the end of each month, all stocks with prices of \$5 or higher are ranked into five quintiles based on the 11-month stock returns (momentum) with a certain lag. Stocks in the top (winners) and bottom (losers) momentum quintiles are further sorted into five portfolios based on uncertainty proxied by the reciprocal of firm age. Stocks are equally weighted and held in the portfolio for 1 month. The first panel depicts the average monthly return differential between the highest- and lowest-uncertainty portfolios for winners and losers, respectively. The second panel depicts the average monthly return differential between winners and losers for the highest- and lowest-uncertainty quintiles, respectively. The broken lines indicate the 95% confidence interval (adjusted for autocorrelation).

much more short-lived following good news than following bad news might explain why high-uncertainty stocks tend to have relatively lower future returns than do low-uncertainty stocks in the overall market (Table II). As shown in the second panel of Figure 1, the return of the momentum strategy is never statistically significant for low-uncertainty stocks, but it is still significant for high-uncertainty stocks even at a lag of 6 months.

Table IX  
Subperiod Analysis

This table summarizes the effect of information uncertainty following bad and good news and its interaction with momentum strategies in two subperiods. The return differential between high- and low-uncertainty stocks is D10 – D1 in Panel A (decile 10 minus decile 1, following the procedure in Table III) and U5–U1 in Panel B (quintile 5 minus quintile 1, following the procedure in Table IV). The interaction with momentum is measured by returns to a zero-investment portfolio with a long position in good-news stocks and a short position in bad-news stocks (GMB) for low- and high-uncertainty, respectively. Firm size (MV) is the market capitalization (in millions of dollars) at the end of month  $t$ . Firm age (AGE) is the number of years since the firm was first covered by CRSP. Analyst coverage (COV) is the number of analysts following the firm in the previous year. Forecast dispersion (DISP) is the standard deviation of analyst forecasts in month  $t$  scaled by the prior year-end stock price. Stock volatility (SIGMA) is the standard deviation of weekly market excess returns over the year ending at the end of month  $t$ . Cash flow volatility (CVOL) is the standard deviation of cash flow from operations in the past 5 years (with a minimum of 3 years), where cash flow from operations is earnings before extraordinary items minus total accruals, scaled by average total assets. Stocks with a price less than \$5 at the portfolio formation date are excluded, and portfolio returns are equally weighted. The sample period is from January 1983 to December 2001;  $t$ -statistics in parentheses are adjusted for autocorrelation.

Panel A: Bad News—Downward Revisions (REV < 0), Good News—Upward Revisions (REV > 0)						
	1/MV	1/AGE	1/COV	DISP	SIGMA	CVOL
Time Period: 1983–1992						
D10 – D1	–0.86%	–0.68%	–0.76%	–0.41%	–1.54%	–1.07%
(bad news)	(–2.18)	(–1.84)	(–2.39)	(–1.09)	(–2.46)	(–2.14)
D10 – D1	0.67%	0.49%	0.62%	–0.20%	0.09%	0.14%
(good news)	(1.65)	(1.23)	(1.23)	(–0.55)	(0.14)	(0.24)
GMB	0.43%	0.54%	0.60%	1.06%	0.60%	0.69%
(low uncertainty)	(2.49)	(2.46)	(2.56)	(4.12)	(4.31)	(3.43)
GMB	1.95%	1.72%	1.98%	1.27%	2.23%	1.91%
(high uncertainty)	(7.25)	(6.44)	(8.64)	(5.04)	(7.00)	(5.93)
Time Period: 1993–2001						
D10 – D1	–0.88%	–1.74%	–0.96%	–0.05%	–1.40%	–0.85%
(bad news)	(–2.07)	(–1.98)	(–1.96)	(–0.10)	(–1.17)	(–1.07)
D10 – D1	1.27%	0.46%	0.67%	1.37%	0.82%	0.96%
(good news)	(2.38)	(0.53)	(1.42)	(2.86)	(0.65)	(1.17)
GMB	0.41%	0.37%	0.69%	0.46%	0.32%	0.75%
(low uncertainty)	(1.63)	(1.45)	(2.19)	(1.22)	(1.78)	(2.93)
GMB	2.56%	2.57%	2.32%	1.87%	2.54%	2.56%
(high uncertainty)	(9.52)	(6.05)	(8.29)	(4.37)	(5.30)	(5.60)
Panel B: Bad News—Past Losers, Good News—Past Winners						
Time Period: 1983–1992						
U5 – U1	–0.34%	–1.20%	–0.81%	–0.75%	–1.37%	–0.76%
(bad news)	(–0.95)	(–3.31)	(–2.22)	(–2.37)	(–3.02)	(–1.97)
U5 – U1	0.75%	0.70%	0.51%	0.64%	0.17%	0.44%
(good news)	(2.41)	(1.93)	(1.69)	(2.51)	(0.37)	(1.15)
GMB	0.49%	0.23%	0.63%	0.81%	0.45%	0.28%
(low uncertainty)	(1.24)	(0.60)	(1.44)	(2.37)	(1.21)	(0.69)
GMB	1.58%	2.14%	1.91%	2.20%	1.98%	1.47%
(high uncertainty)	(5.04)	(6.10)	(5.76)	(6.79)	(5.00)	(4.13)

(continued)

Table IX—Continued

Panel B: Bad News—Past Losers, Good News—Past Winners						
	1/MV	1/AGE	1/COV	DISP	SIGMA	CVOL
Time Period: 1993–2001						
U5 – U1	–0.47%	–2.19%	–0.98%	–0.22%	–1.58%	–1.22%
(bad news)	(–0.90)	(–3.20)	(–2.05)	(–0.47)	(–1.77)	(–2.19)
U5 – U1	1.38%	0.96%	0.25%	0.73%	0.95%	0.35%
(good news)	(3.15)	(1.84)	(0.56)	(2.10)	(0.94)	(0.58)
GMB	1.10%	0.59%	1.23%	1.46%	0.84%	1.13%
(low uncertainty)	(1.31)	(0.90)	(1.61)	(1.79)	(1.64)	(1.79)
GMB	2.95%	3.75%	2.45%	2.41%	3.37%	2.70%
(high uncertainty)	(4.64)	(5.06)	(3.62)	(3.85)	(4.09)	(3.69)

C. Subperiod Analysis

In Table IX, I check the robustness of the results across time periods to see if they are time-specific. This analysis will also show if investor behavior changes over time. Arguably, the information environment has become richer and investors might learn from past mistakes. In Table IX, I report results for the 1983 to 1992 and 1993 to 2001 subperiods. The return differentials between high- and low-uncertainty stocks following good or bad news are similar in these two subperiods. Although the uncertainty effect is insignificant in some bad- or good-news cases, the effect of information uncertainty on momentum trading strategies is still evident. In both subperiods, a zero-investment portfolio with a long position in good-news stocks and a short position in bad-news stocks generates much higher returns for high-uncertainty portfolios than it does for low-uncertainty portfolios. Overall, the return patterns are similar in these two subperiods, although the later subperiod has more firms with good news due to the booming economy from 1992 to 1999.

D. Analysis on NYSE Stocks Only

To ensure that the results are not driven by a few small stocks, I also check the robustness of the results using only NYSE stocks. The returns (untabulated) are consistent with previous results. In fact, the uncertainty effect as measured by the return differential between high- and low-uncertainty stocks is more pronounced following good news than following bad news, which complements the evidence from the whole sample in supporting my hypothesis. Although this result may be partly due to the fact that non-NYSE stocks have more short-sale restrictions, I am otherwise unable to explain this feature of the data.

I also conduct other robustness checks for the whole sample, such as holding a stock in the portfolio for longer than 1 month, with a portion of each portfolio rebalanced monthly. The return follows a similar pattern to that previously observed. Finally, I try independent sorts or different sorting orders in portfolio

formation, such as sorting stocks first by information uncertainty proxy and then by momentum. The results are robust to these tests.

## VII. Conclusion

In this paper, I examine the role of information uncertainty in short-term price continuation anomalies and cross-sectional variations in stock returns. I use analyst forecast revisions and price momentum to distinguish good news from bad news and use firm size, firm age, analyst coverage, dispersion in analyst earnings forecasts, stock volatility, and cash flow volatility to proxy for information uncertainty.

There is clear evidence that the initial market reaction to new public information is incomplete, which implies that bad news predicts relatively lower future returns and good news predicts relatively higher future returns. More importantly, the degree of incompleteness of the market reaction increases monotonically with the level of information uncertainty, suggesting that investors tend to underreact more to new information when there is more ambiguity with respect to its implications for firm value. As a result, greater information uncertainty produces relatively lower future returns following bad news and relatively higher future returns following good news. The opposite effects of information uncertainty on stock returns following good versus bad news amplify the profitability of certain trading strategies. For example, the momentum strategy works particularly well when limited to high-uncertainty stocks.

Although I cannot definitively rule out the possibility that each information uncertainty proxy may capture other effects, the six proxies draw a consistent picture that investors underreact to a higher degree when there is greater information uncertainty. The predictability of stock returns based on the nature of news and the level of uncertainty is of its own value to individual investors and fund managers. For researchers and standard setters, there are more fundamental questions to be addressed. The evidence that greater information uncertainty predicts higher expected returns following good news and lower expected returns following bad news is inconsistent with the notion that information uncertainty is a cross-sectional risk factor and is compensated by higher stock returns. It also suggests that price and earnings momentum are more likely to be rooted in a failure of information to flow completely into stock prices.

## REFERENCES

- Ang, Andrew, Robert Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2003, The cross-section of volatility and expected returns, Working paper, Columbia University.
- Baber, William R., and Sok-Hyon Kang, 2002, The impact of split adjusting and rounding on analysts' forecast error calculations, *Accounting Horizon* 16, 277–290.
- Barberis, Nicholas, Andrei Shleifer, and Robert Vishny, 1998, A model of investor sentiment, *Journal of Financial Economics* 49, 307–343.

- Barron, Orie, Oliver Kim, Steve Lim, and Douglas Stevens, 1998, Using analysts' forecasts to measure properties of analysts' information environment, *The Accounting Review* 73, 421–433.
- Barron, Orie, and Pamela Stuerke, 1998, Dispersion in analyst's earnings forecasts as a measure of uncertainty, *Journal of Accounting, Auditing, and Finance* 13, 243–268.
- Barry, Christopher B., and Stephen J. Brown, 1985, Differential information and security market equilibrium, *Journal of Financial and Quantitative Analysis* 20, 407–422.
- Bernard, Victor, and Jacob Thomas, 1990, Evidence that stock prices do not fully reflect the implications of current earnings for future earnings, *Journal of Accounting & Economics* 13, 305–340.
- Botosan, Christine, 1997, Disclosure level and the cost of equity capital, *The Accounting Review* 72, 323–349.
- Botosan, Christine, and Marlene Plumlee, 2002, A re-examination of disclosure level and the expected cost of equity capital, *Journal of Accounting Research* 40, 21–40.
- Carhart, Mark, 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.
- Chan, Louis K.C., Narasimhan Jegadeesh, and Josef Lakonishok, 1996, Momentum strategies, *Journal of Finance* 51, 1681–1713.
- Chan, Louis K.C., Narasimhan Jegadeesh, and Josef Lakonishok, 1999, The profitability of momentum strategies, *Financial Analysts Journal*, 80–90.
- Cohen, Daniel, 2003, Quality of financial reporting choice: Determinants and economic consequences, Working paper, Northwestern University.
- Coles, Jeffrey, and Uri Loewenstein, 1988, Equilibrium pricing and portfolio composition in the presence of uncertain parameters, *Journal of Financial Economics*, 279–303.
- Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam, 1998, Investor psychology and security market over- and under-reactions, *Journal of Finance* 53, 1839–1886.
- Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam, 2001, Overconfidence, arbitrage, and equilibrium asset pricing, *Journal of Finance* 56, 921–965.
- Daniel, Kent, and Sheridan Titman, 1999, Market efficiency in an irrational world, *Financial Analysts' Journal* 55, 28–40.
- Diamond, Douglas, and Robert Verrecchia, 1991, Disclosure, liquidity and the cost of equity capital, *The Journal of Finance* 46, 1325–1360.
- Diether, Karl. B., Christopher J. Malloy, and Anna Scherbina, 2002, Difference of opinion and the cross section of stock returns, *The Journal of Finance* 57, 2113–2141.
- Easley, David, and Maureen O'Hara, 2001, Information and the cost of capital, Working paper, Cornell University.
- Fama, Eugene F., 1998, Market efficiency, long-term returns, and behavioral finance, *Journal of Financial Economics* 49, 283–306.
- Fama, Eugene F., and Kenneth R. French, 1995, Size and book-to-market factors in earnings and returns, *The Journal of Finance* 50, 131–155.
- Fama, Eugene F., and Kenneth R. French, 1996, Multifactor explanations of asset pricing anomalies, *The Journal of Finance* 51, 55–84.
- Gleason, Cristi A., and Charles M.C. Lee, 2003, Analyst forecast revisions and market price discovery, *The Accounting Review* 78, 193–225.
- Hansen, Lars, and Ravi Jagannathan, 1991, Implications of security market data for models of dynamic economics, *Journal of Political Economy* 99, 225–262.
- Hirshleifer, David, 2001, Investor psychology and asset pricing, *Journal of Finance* 56, 1533–1596.
- Hong, Harrison, Terry Lim, and James C. Stein, 2000, Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies, *Journal of Finance* 55, 265–295.
- Hribar, Paul, and Daniel Collins, 2002, Errors in estimating accruals: Implications for empirical research, *Journal of Accounting Research* 40, 105–134.
- Imhoff, Eugene, and Gerald Lobo, 1992, The effect of ex ante earnings uncertainty on earnings response coefficients, *The Accounting Review* 67, 427–439.
- Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *The Journal of Finance* 56, 699–720.

Copyright of Journal of Finance is the property of Blackwell Publishing Limited and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.



# Exhibit 63

No. 03-11064  
United States Court of Appeals, Fifth Circuit

## Barrie v. Intervice-Brite, Inc.

409 F.3d 653 (5th Cir. 2005)  
Decided May 12, 2005

No. 03-11064.

655 May 12, 2005. \*655

Susan K. Alexander (argued), Lerach, Coughlin, Stoia, Geller, Rudman Robbins, LLP, San Francisco, CA, Roger L. Mandel, Stanley, Mandel Iola, Dallas, TX, for Plaintiffs-Appellants.

G. Luke Ashley (argued), Timothy R. McCormick, Richard Barrett Phillips, Thompson Knight, Dallas, TX, for Defendants-Appellees.

Appeal from the United States District Court for the Northern District of Texas.

### ORDER ON REHEARING

(Opinion Jan. 12, 2005, 5th Cir., [397 F.3d 249](#))

Before BENAVIDES, DENNIS and CLEMENT,  
Circuit Judges.

---

EDITH BROWN CLEMENT, Circuit Judge:

A. The panel's opinion issued on January 12, 2005 inadvertently made conflicting rulings regarding the allegations contained in paragraphs 38, 59, and 68 of the plaintiffs' complaint. Having duly considered the petition for rehearing, response, and reply, and the appellees' letter submitted pursuant to [FED. R.APP. P. 28\(j\)](#) and the appellants' response thereto, we clarify our ruling by hereby modifying the opinion in the following respects as set out in (1) and (2) below, viz:

(1) The final paragraph of Section III.A.(2)a, which begins with the phrase "Six of the statements" is hereby deleted from the opinion and

replaced by the following:

Three of the statements fail to adequately identify the speaker. Complaint at ¶¶ 29, 44, and 56. Paragraphs 44 and 56 are press releases and contain no specific allegation as to who made the statement. *Id.* at ¶¶ 44, 56. In *Southland*, this Court affirmed the proposition that

the PSLRA requires the plaintiffs to "distinguish among those they sue and enlighten *each defendant* as to his or her particular part in the alleged fraud." As such, corporate officers may not be held responsible for unattributed corporate statements solely on the basis of their titles, even if their general level of day-to-day involvement in the corporation's affairs is pleaded.

365 F.3d at 365 (quoting *Southland Sec. Corp. v. INSpire Ins. Solutions, Inc.*, No. 4:00-CV-355-Y, 2002 WL 32453742, at \*6, 2002 U.S. Dist. LEXIS 26659, at \*20-21 (N.D. Tex. April 2, 2002)). Although this Court stated that "corporate documents that have no stated author or statements within documents not attributed to any individual may be charged to one or more corporate officers provided specific factual allegations link the individual to the statement at issue," *id.*, there are no such factual allegations here linking the individual defendants to the statements in the press releases. Consequently, these allegations were properly dismissed. Paragraph 29 attributes false and misleading statements to "management." Complaint at ¶ 29. Again, this fails pursuant to *Southland's* rejection of the group pleading doctrine.

Paragraphs 38, 59, and 68 attribute statements to "Hammond and Graham." *Id.* at ¶¶ 38, 59, and 68. The district court held that "the allegation that 'Hammond and Graham' made a statement does not identify the speaker; it merely narrows the range of possible speakers down to two people." The plaintiffs argue, however, that defendants who silently listened as others made statements that they knew were false are liable for their omission in failing to correct a falsehood. They claim that whether a particular statement during a conference call or a road show was uttered by Hammond or Graham, both are liable: one for the utterance, and the \*656 other for the omission in failing to correct the falsehood.

As one district court held:

a high ranking company official cannot sit quietly at a conference with analysts, knowing that another official is making false statements and hope to escape liability for those statements. If nothing else, the former official is at fault for a material omission in failing to correct such statements in that context.

*In re SmarTalk Teleservices, Inc. Sec. Litig.*, 124 F.Supp.2d 527, 543 (D.Ohio 2000).

The defendants argue that this reasoning is inapplicable in this case, because the Complaint did not specify whether Hammond spoke and Graham failed to correct, or vice versa. They assert that the allegations concerning Hammond and Graham are deficient given this Court's rejection of the group pleading doctrine in *Southland*.

656

We reject the defendants' argument. Where it is pled that one defendant knowingly uttered a false statement and the other defendant knowingly failed to correct it, even if it is not alleged which defendant made the statement and which defendant did not correct it, the fraud is sufficiently pleaded as to each defendant. This is not inconsistent with the plain language of subsection (1) of 15 U.S.C. § 78u-4(b), and accords with common sense and the policy considerations underlying the heightened pleading requirements. The *Southland* Court explained: "In securities fraud suits, this heightened pleading standard provides defendants with fair notice of the plaintiffs' claims, protects defendants from harm to their reputation and goodwill, reduces the number of strike suits, and prevents plaintiffs from filing baseless claims and then attempting to discover unknown wrongs." 365 F.3d at 363 (quoting *Tuchman v. DSC Communications*, 14 F.3d 1061, 1067 (5th Cir. 1994)). The Complaint's allegations that false statements were made by Hammond and Graham imply that one of them made the statement and the other knowingly failed to correct it. Accordingly, both Hammond and Graham are on fair notice of the claims against them. The district court erred in dismissing the claims attributed to Hammond and Graham when the allegations sufficiently indicated that

one had spoken the fraudulent statement, and the other had failed to correct it.

By the same reasoning, Smith is properly liable for his alleged omission to correct Hammond and Graham's alleged misrepresentation in the conference call described in paragraph 38 of the Complaint. Although the Complaint does not specify whether it was Hammond or Graham who uttered the alleged falsehood, it does contend with sufficient particularity that Smith failed to correct it. Smith is also properly liable for his alleged omission to correct Hammond's alleged misrepresentation in the conference call described in paragraph 58 of the Complaint.

(2) The final five paragraphs of section Section III.A.(2)b, which begin with the sentence "The plaintiffs also argue that defendants who silently listened as others made statements that they knew were false are liable for their omission in failing to correct a falsehood," are deleted from the opinion.

B. Except as above provided, the panel opinion issued herein January 12, 2005 is unchanged.

657 C. The Petition for Rehearing is DENIED. \*657

# Exhibit 64

No. 12-1750

United States Court of Appeals, First Circuit.

## Bricklayers & Trowel Trades International Pension Fund v. Credit Suisse Securities (USA) LLC

752 F.3d 82 (1st Cir. 2014)

Decided May 14, 2014

No. 12–1750.

2014-05-14

BRICKLAYERS AND TROWEL TRADES INTERNATIONAL PENSION FUND, Plaintiff, Appellant, James Uphoff; Goodman Family Trust; Malka Birnbaum, on behalf of herself and all others similarly situated; Neil McCarty; Rodney W. Narbesky, individually and on behalf of all others similarly situated, Plaintiffs, v. CREDIT SUISSE SECURITIES (USA) LLC; Credit Suisse (USA), Inc.; Jamie Kiggen; Frank P. Quattrone; Laura Martin; Elliot Rogers, Defendants, Appellees.

Frederic S. Fox, with whom Kaplan Fox & Kilsheimer LLP and Shapiro Haber & Urmy LLP were on brief, for appellant. Lawrence Portnoy, with whom Daniel J. Schwartz, Jonathan K. Chang, Dharma Betancourt Frederick, Davis Polk & Wardwell LLP, Robert Buhlman, Siobhan E. Mee, Amanda V. Muller, and Bingham McCutchen LLP were on brief, for appellees.

HOWARD

84 \*84

Frederic S. Fox, with whom Kaplan Fox & Kilsheimer LLP and Shapiro Haber & Urmy LLP were on brief, for appellant. Lawrence Portnoy, with whom Daniel J. Schwartz, Jonathan K. Chang, Dharma Betancourt Frederick, Davis Polk & Wardwell LLP, Robert Buhlman, Siobhan E. Mee, Amanda V. Muller, and Bingham McCutchen LLP were on brief, for appellees.

**Before HOWARD, Circuit Judge, SOUTER,\* Associate Justice and TORRESEN,\*\* District Judge.**

**HOWARD, Circuit Judge.**

\* Hon. David H. Souter, Associate Justice (Ret.) of the Supreme Court of the United States, sitting by designation.

\*\* Of the District of Maine, sitting by designation.

Alleging violations of Sections 10(b) and 20(a) of the Securities Exchange Act and of SEC Rule 10b5, the appellant pension fund and other America Online (“AOL”) shareholders brought this class action against Credit Suisse First Boston (“CSFB”), former CSFB analysts Jamie Kiggen and Laura Martin, and other related defendants. The shareholders claim that CSFB fraudulently withheld relevant information from the market in its reporting on the AOL–Time Warner merger, and that the shareholders purchased stock in the new company at prices that were artificially inflated as a result of the defendants' purposeful omissions. This appeal concerns the admissibility of the opinion of the shareholders' expert Dr. Scott D. Hakala, whose testimony the district court precluded for lack of reliability. We find no abuse of discretion in that decision. We also agree with the district court that, without the expert's testimony, the shareholders are unable to establish loss causation. Summary judgment was therefore properly awarded to the defendants.

## I. Background

### A. Facts

On January 11, 2001, Time Warner Inc. and AOL merged into a single media and technology company (hereinafter referred to as “AOL”). This marriage of “old” and “new” media received extensive coverage from both the press and the financial industry. CSFB was among the many financial firms reporting on AOL’s business and forecasting its outlook for the future. Kiggen and Martin headed CSFB’s AOL coverage beginning the day after the merger and continuing for about a year, through January 2002, when CSFB ceased covering AOL (Kiggen retired in January 2002; Martin had left CSFB a few months earlier). During the coverage period, CSFB published the results of its research in regular reports. These contained, in addition to observations about AOL, a buy or sell recommendation and a price target, 85 which was a prediction of AOL’s stock \*85 price twelve months hence. CSFB issued thirty-five such reports during this period, and each such report recommended buying AOL stock. CSFB initially targeted AOL’s future stock price at \$80, but revised it downwardly to \$75 one month later in February 2001, and then to \$45 in September 2001. Nine months later, AOL’s stock was trading at \$11 per share.

The shareholders allege that Kiggen and Martin misrepresented their true opinions in these reports, in order to maintain a good relationship with AOL. The shareholders’ theory is that AOL had the potential to generate significant investment banking revenue for CSFB, and Kiggen and Martin overstated AOL’s financial strength in the hopes of winning this future business (CSFB did in fact assist AOL in managing a bond deal purportedly generating between \$750,000 and \$820,000 in fees for CSFB). In a series of internal emails among AOL team members, Kiggen and Martin expressed doubts about their projections for AOL, yet decided not to lower their estimates for AOL’s future performance notwithstanding these concerns. Moreover, they regularly showed

their projections to AOL and revised them based on AOL’s reactions. Even as advertising revenue, a key factor in AOL’s success, declined throughout the industry, CSFB reports continued to predict AOL’s ability to rise above the general slowdown.

In addition, the shareholders allege two instances in which CSFB<sup>1</sup> received non-public, material information about AOL that CSFB did not disclose in its coverage of the company. On July 10 and 11, 2001, Anthony Lorenzo, a junior CSFB analyst not assigned to cover AOL, emailed to Kiggen information about AOL layoffs. Citing an unnamed source, Lorenzo wrote that AOL “apparently ... had some layoffs” that “were medium in terms of severity and will not be announced publicly.” The parties disagree over the import of this tip. The shareholders claim that the information pertained to layoffs of “up to 1,000 employees” subsequently reported in *The Wall Street Journal* and *The Washington Post* on August 13 and 14, 2001. CSFB counters that this unnamed source (later identified as a low-level employee in AOL’s Interactive Marketing Group) was referring only to a small number of layoffs that occurred within the Interactive Marketing Group on July 10, 2001, as reported in *The Washington Post* the next day.

<sup>1</sup> At times, we refer to the defendants, collectively, as “CSFB”.

Lorenzo’s emails also mentioned “that AOL was under investigation and has suspended some employees for inappropriate accounting activities—some deals booked inappropriately inflated revenue.” CSFB did not disclose this information in any of its reports; it was eventually reported by *The Washington Post* in a July 2002 article disclosing that AOL had engaged in “unconventional” advertising deals that might have inflated revenue. On July 24, 2002, AOL



acknowledged that the SEC was investigating its accounting practices, but denied any wrongdoing.

### B. Procedural History

On the basis of these alleged material misstatements and omissions—overstating AOL's financial strength, not disclosing reports of medium-severity layoffs, and not disclosing reports of unconventional accounting—the shareholders brought suit in December 2005 against Kiggen, Martin, and CSFB under Section 10(b) of the Exchange Act, 15 U.S.C. § 78j(b), and under SEC Rule 10b–5. The complaint also alleged that CSFB, Credit Suisse First Boston (USA) Inc. (CSFB's parent company), and CSFB executives Frank Quattrone and Elliot Rogers  
86 violated Section 20(a) of \*86 the Exchange Act, 15 U.S.C. § 78t(a), by failing to exercise control over their employees' alleged misstatements and omissions.

In due course, the defendants sought summary judgment. At the hearing occasioned by that motion, the shareholders and the defendants each presented expert testimony to show the effect, or lack thereof, of CSFB's omissions on AOL stock prices. The shareholders retained Dr. Hakala, while CSFB employed Dr. René M. Stulz. Each side subsequently moved to exclude the other's expert opinion under *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579, 597, 113 S.Ct. 2786, 125 L.Ed.2d 469 (1993) (“[T]he Rules of Evidence—especially Rule 702—[ ] assign to the trial judge the task of ensuring that an expert's testimony both rests on a reliable foundation and is relevant to the task at hand.”). In due course, the court held a *Daubert* hearing to determine the admissibility of the proffered expert testimony on loss causation. C. Event Studies and Expert Testimony

Loss causation is among the six elements of a private cause of action for securities fraud; the other five are: a material misrepresentation or omission, scienter, a connection with the purchase or sale of a security, reliance, and economic loss.

*Dura Pharm., Inc. v. Broudo*, 544 U.S. 336, 341–42, 125 S.Ct. 1627, 161 L.Ed.2d 577 (2005). To prove loss causation, a plaintiff “must show ‘a sufficient connection between [the fraudulent conduct] and the losses suffered....’ ” *In re Omnicom Grp., Inc. Sec. Litig.*, 597 F.3d 501, 510 (2d Cir.2010) (quoting *Lattanzio v. Deloitte & Touche LLP*, 476 F.3d 147, 157 (2d Cir.2007)) (alterations in original). In other words, the stock market must have reacted to the subsequent disclosure of the misconduct and not to a “tangle of [other] factors.” *Dura Pharm.*, 544 U.S. at 343, 125 S.Ct. 1627.

The usual—it is fair to say “preferred”—method of proving loss causation in a securities fraud case is through an event study, in which an expert determines the extent to which the changes in the price of a security result from events such as disclosure of negative information about a company, and the extent to which those changes result from other factors.<sup>2</sup> First, the expert selects the period in which the event could have affected the market price.<sup>3</sup> The expert then attempts to determine the effect on the share price of general market conditions, as opposed to company-specific events, using a multiple regression analysis, a statistical means for explaining the relationship between two or more variables. 1 David L. Faigman et al., *Modern Scientific Evidence; The Law and Science of Expert Testimony* 430 (2012). Thus, for any given day, the expert predicts the company's share price based on the market trends on that particular day. The expert then compares this predicted return with the actual return in the event window in order to determine the probability that an abnormal return of that magnitude could have occurred by chance. If this probability is small enough, the expert can  
87 reject the hypothesis\*87 that normal market fluctuations, as opposed to company-specific events, can explain the movement in the share price.

<sup>2</sup> For additional information about event studies in litigation, see Sanjai Bhagat & Roberta Romano, *Event Studies and the Law: Part I: Technique and Corporate Litigation*, 4 Am. L. & Econ. Rev. 141 (2002), and Michael J. Kaufman & John M. Wunderlich, *Regressing: The Troubling Dispositive Role of Event Studies in Securities Fraud Litigation*, 15 Stan. J.L. Bus. & Fin. 183, 186 (2009).

<sup>3</sup> “Stock price,” “share price,” “market price,” “closing price,” and “return” are all used interchangeably throughout this opinion.

Central to multiple regression analyses are variables, which, as the term implies, can have two or more possible values. *Id.* n. 1. Multiple regression includes a variable to be explained (the dependent variable) and explanatory (or independent) variables that have the potential to be associated with changes to the dependent variable. *Id.* at 430. (“[A] multiple regression analysis might estimate the effect of the number of years of work on salary. Salary would be the dependent variable to be explained; years of experience would be the explanatory variable.”). The third type of variable at issue in this case is a dummy variable, which is also known as a “binary variable” because it only has two possible values, such as gender, or, as in this case, the existence or non-existence of company-specific events.<sup>4</sup> By assigning the variable a value of zero or one in the mathematical formulae used in the analysis, the dummy variable becomes mutually exclusive with respect to any explanatory variables, unable to exist or affect the outcome simultaneously. Thus, by using a dummy variable, the projected various outcomes can reveal which explanatory variables affect the dependent variable. **D. Dr. Hakala's Event Study**

<sup>4</sup> An opinion from the District of New Jersey provides a succinct example of the use of a dummy variable:

[S]uppose you are investigating [United States] consumption behavior with time series data for the period 1930 to 1950. You would expect that consumption behavior would have been significantly different during the years of World War II than it was before and after the war. To take this effect into account, you can create an artificial variable that will take the value 1 during each of the war years and the value 0 during each of the other years.

*Animal Sci. Prods., Inc. v. China Nat'l Metals & Minerals Imp. & Exp. Corp.*, 702 F.Supp.2d 320, 358 n. 44 (D.N.J.2010).

Such is the basic structure of an event study. In its motion to exclude Dr. Hakala's testimony, CSFB alleged that his methodology included techniques that did not meet the standards of reliability articulated in *Daubert*. It challenged four elements of Dr. Hakala's study.

### 1. Selection of Event Dates

The first alleged flaw in Dr. Hakala's analysis was his selection of event dates. CSFB claimed that Dr. Hakala failed to conform to event study methodology by selecting his event dates *after* running his regression analysis. As noted previously, the first step of an event study is identifying the relevant dates that are the focus of the study. CSFB argued that Dr. Hakala reversed the steps in this process, first conducting a regression analysis, and then, after identifying fifty-seven dates with statistically significant abnormal returns, using them as the relevant dates for his event study.

According to CSFB, this results-driven approach produced event dates that had “little relationship with the allegations or facts in this case and ma[de] no sense even under [Dr. Hakala's] own definition of ‘relevance.’ ” For instance, Dr. Hakala attributed some abnormal market increases

in AOL stock prices to the defendants on days when CSFB released no reports about AOL—often when it was no longer reporting about AOL. Dr. Hakala also characterized several of the dates in his study as corrective, despite the fact that the complaint had labeled them as inflationary.<sup>5</sup> On

88 one event date, the abnormal \*88 market return was negative, yet Dr. Hakala classified the date as inflationary. Finally, Dr. Hakala often identified dates as corrective when no negative information entered the market, and other dates as inflationary when no positive information entered the market.

<sup>5</sup> An inflationary date occurred when misinformation or omissions inflated AOL's stock price. A corrective date occurred when truthful information caused AOL's stock price to return to its normal levels.

## 2. Overuse of Dummy Variables

CSFB also asserted that Dr. Hakala's use of dummy variables not only overstated the baseline stability of AOL's stock prices, but also failed to satisfy the *Daubert* requirement of reproducibility. The goal of a regression analysis is to create a baseline against which the market return on event dates is measured. Through the use of dummy variables, the event dates themselves are excluded from, or “dummied out” of, the regression analysis to indicate the presence or absence of some event. This is designed to prevent the event days themselves from distorting the baseline. Dr. Hakala, in addition to dummied out relevant event dates, dummied out all dates containing material news about AOL.<sup>6</sup> He chose this approach to control for days when AOL's stock price might have fluctuated due to the release of information that was, for purposes of this litigation, irrelevant. He believed that these material news dates could improperly influence the baseline regression, and cited other financial economists who endorse this methodology. Using this approach, Dr. Hakala dummied out 211 out of

388 days in the study period—54% of the total number of days. CSFB argued that Dr. Hakala's approach went too far, creating an unrealistically stable baseline and thereby ensuring that all relevant event dates would appear more unusual than they really were.

<sup>6</sup> The terms “relevant event dates” and “material news dates,” though similar, are distinct. Relevant event dates are the fifty-seven dates that are the focus of Dr. Hakala's event study. Material news dates refer to the additional one hundred fifty-four dates Dr. Hakala dummied out of his event study because they contained material news.

CSFB also attacked Dr. Hakala's dummy selection as arbitrary. Dr. Hakala performed three event studies relating to the America Online–Time Warner merger. Although he used the same criteria to select the material dates each time,<sup>7</sup> he dummied out more material dates in each subsequent study. CSFB argued that his selection criteria were so vague that two economists would be apt to pick vastly different numbers of material dates given the same instructions. Thus, CSFB argued, Dr. Hakala's methodology cannot be replicated. *See Daubert*, 509 U.S. at 593, 113 S.Ct. 2786 (“Ordinarily, a key question to be answered in determining whether a theory or technique is scientific knowledge that will assist the trier of fact will be whether it can be (and has been) tested.”). The proof of this flaw, according to the defendants, was the fact that even Dr. Hakala could not select the same number of material news dates in three separate event studies.

<sup>7</sup> Dr. Hakala's criteria came from “the NASDAQ guidelines as recognized by the SEC.” *See* Self-Regulatory Organizations; Notice of Filing of Proposed Rule Change by the National Association of Securities Dealers, Inc. Relating to Issuer Disclosure of Material Information, 67 F.R. 51,306 (Jul. 31, 2002). He also included “third

party news and reports, and analysts' reports to that list consistent with the academic studies.”

### 3. Previously Disclosed Information

In *Basic Inc. v. Levinson*, 485 U.S. 224, 108 S.Ct. 978, 99 L.Ed.2d 194 (1988), the Supreme Court held that a plaintiff in a securities fraud suit need  
89 not prove individual reliance on the defendants’<sup>89</sup> fraudulent statements when purchasing company stock. *Id.* at 247, 108 S.Ct. 978. Instead, courts will presume reliance as long as the company’s shares trade in an efficient market, that is, one which incorporates all public statements about the company—including the defendants’ fraudulent statements—into its share price. *Id.* “An investor who buys or sells stock at the price set by the market does so in reliance on the integrity of that price.” *Id.* Consequently, investors must also implicitly rely on the integrity of the information affecting the stock price. Investors who avail themselves of the fraud-on-the-market theory recognized in *Basic*, however, must be consistent. If it is assumed that the market reacts to the fraud, it must also be assumed that it reacts to the truth. Accordingly, once a misstatement or corrective disclosure is publicly known in an efficient market, courts will assume that the stock price reacts immediately, and any claim that an event moved the stock price when the event was not actually a new disclosure will necessarily fail.<sup>8</sup>

<sup>8</sup> A case pending before the Supreme Court has raised the issue of the continuing viability of the fraud-on-the-market theory. See *Halliburton Co. v. Erica P. John Fund, Inc.*, No. 13–317 (U.S. argued March 5, 2014).

CSFB argued that Dr. Hakala’s event study included some “new disclosures” that were not in fact new to the market. CSFB pointed to several instances in Dr. Hakala’s event study when he attributed the rise or fall of AOL’s stock price to

the disclosure of “stale” information. Consequently, CSFB averred, this information could not form the basis of a proper event date, and Dr. Hakala’s rejection of the efficient market hypothesis rendered his study inadmissible.

### 4. Failure to Control for Confounding Factors

The purpose of an event study, as noted, is to isolate the impact of an alleged misstatement, omission, or disclosure on the stock price. A recurring problem in event studies is the presence of “confounding factors”—news stories, statements, or events that coincide with relevant event dates and that themselves potentially affect the company’s stock price. CSFB claimed that Dr. Hakala made no attempt to control for the many confounding news stories that emerged at the same time as CSFB reports and other relevant events, and therefore that his event study did not show that CSFB’s statements, as opposed to some other news story, moved the stock price on any given day. **E. The District Court’s Opinion**

In January 2012, the district court issued an order precluding Dr. Hakala’s testimony, relying on the four factors argued in CSFB’s motion to preclude. While it gave specific examples for each factor, the court explained that these were illustrative of pervasive problems.

With respect to the event date selection, the district court determined that

“[r]ather than study the market’s reaction to the misrepresentations alleged in the complaint, Dr. Hakala cherry-picked unusually volatile days and made them the focus of his study. If the stock price increased sharply, he attributed it to the defendants (even if no CSFB reports were released on that day). If the stock price decreased sharply, he called it a corrective disclosure (even if the news released was positive). The Court concludes ... that, quite simply, Dr. Hakala’s theory does not match the facts.

<sup>90</sup> \*90 *Bricklayers & Trowel Trades Int’l Pension*



*Fund v. Credit Suisse First Boston*, 853 F.Supp.2d 181, 188 (D.Mass.2012) (citations omitted) (internal quotation marks and alterations omitted), reconsideration denied, published in 853 F.Supp.2d 181, 195 (D.Mass. May 17, 2012).

Next, and with little independent analysis, the district court followed the reasoning of two other district courts in concluding that Dr. Hakala's use of dummy variables was also unreliable. See *In re Northfield Labs., Inc. Sec. Litig.*, 267 F.R.D. 536, 548 (N.D.Ill.2010); *In re Xcelera.com Sec. Litig.*, No. 00-11649-RWZ, 2008 WL 7084626, at \*1 (D.Mass. Apr. 25, 2008). In those cases, courts had criticized the high percentage of dummied-out dates in Dr. Hakala's studies, finding that the practice artificially stabilized the baseline regression. Here, the district court noted that Dr. Hakala had dummied out a higher percentage of days in this study than in either of those cases. It concluded that “[i]f those courts were correct in excluding his event studies ..., as this Court believes they were, it follows *a fortiori* that his event study should be excluded here.” *Bricklayers*, 853 F.Supp.2d at 189.

The district court also found that Dr. Hakala's study “repeatedly ignores the efficient market principle” by attributing price fluctuations to previously disclosed information. *Id.* The shareholders' attempt to “presume an efficient market to prove reliance and an inefficient market to prove loss causation,” according to the district court, was tantamount to “hav[ing] their cake and eat[ing] it too.” *Id.* at 190.

Finally, the district court rejected Dr. Hakala's attempt to disaggregate confounding information on the event dates. It acknowledged that AOL received near uninterrupted coverage during the Class Period, making Dr. Hakala's task difficult. But it concluded that Dr. Hakala did not use an accepted means for separating the impact of the relevant event from the impact of confounding information. As an example of one method that Dr. Hakala could have used, the district court

discussed intra-day trading analysis. This analysis requires tracking the stock price throughout the day to see whether its daily highs or lows correspond with the relevant event, or with the release of some other information. Dr. Hakala did not do this. Instead, he “either attributed a rough proportion of the movement to each report or blamed it all on the defendants.” *Id.* at 191. The district court considered this approach unreliable and unscientific.

The court ultimately concluded that Dr. Hakala's event study lacked sufficient reliability to be presented to a jury. It indicated that “[h]ad Dr. Hakala's event study suffered from only one of the four methodological defects identified by this Court, or suffered from those flaws jointly but to a lesser degree, today's ruling might have been different,” *id.* at 191, but, given the extent of Dr. Hakala's errors, preclusion was necessary.

The district court awarded summary judgment to CSFB *sua sponte*, deciding that Dr. Hakala's event study, “even if it were admitted,” did not raise a triable issue of loss causation.<sup>9</sup> *Id.* at 191–92. The district court's reasoning largely restated the problems that persuaded it to preclude Dr. Hakala's event study.

<sup>9</sup> The defendants' original motion for summary judgment had been denied earlier, subject to being revisited if the court determined that Dr. Hakala's testimony should be excluded.

After their motion for reconsideration was denied, the shareholders appealed both the preclusion of Dr. Hakala's event study and the grant of summary judgment.\*91

## II. Analysis

### A. Expert Testimony

We review a district court's decision to exclude an expert witness's testimony for abuse of discretion. *Milward v. Acuity Specialty Prods. Grp., Inc.*, 639

F.3d 11, 13 (1st Cir.2011). “This standard is not monolithic: within it, embedded findings of fact are reviewed for clear error, questions of law are reviewed de novo, and judgment calls are subjected to classic abuse-of-discretion review.” *Ungar v. Palestine Liberation Org.*, 599 F.3d 79, 83 (1st Cir.2010).

Since the Supreme Court's decision in *Daubert*, trial judges have acted as gatekeepers of expert testimony, assessing it for reliability before admitting it. See *Milward*, 639 F.3d at 14. Expert testimony comes in many different forms, but certain non-exclusive factors can assist a trial court in its task: “(1) whether the theory or technique can be and has been tested; (2) whether the technique has been subject to peer review and publication; (3) the technique's known or potential rate of error; and (4) the level of the theory or technique's acceptance within the relevant discipline.” *United States v. Mooney*, 315 F.3d 54, 62 (1st Cir.2002) (citing *Daubert*, 509 U.S. at 593–94, 113 S.Ct. 2786). Moreover, an expert's opinion must be relevant “not only in the sense that all evidence must be relevant, but also in the incremental sense that the expert's proposed opinion, if admitted, likely would assist the trier of fact to understand or determine a fact in issue.” *Ruiz–Troche v. Pepsi Cola of P.R. Bottling Co.*, 161 F.3d 77, 81 (1st Cir.1998) (citations omitted).

As the district court observed, no single factor is dispositive in determining the admissibility of Dr. Hakala's expert testimony. Consequently, we will address the four factors from the district court's opinion individually before analyzing the overall admissibility of Dr. Hakala's testimony

### 1. Selection of Event Dates

The district court committed no abuse of discretion in concluding that Dr. Hakala selected event dates based on unreliable criteria. Event selection should not be difficult to understand, yet Dr. Hakala's event study leaves us guessing as to how he chose the fifty-seven dates included in his study. He certainly did not rely on the

shareholders' complaint. Not only did Dr. Hakala include many dates that bear no relationship to the allegations in the complaint, in some instances he has turned the complaint on its head, treating certain events as corrective when the complaint labeled them inflationary. This complete disconnect between the event study and the complaint nullifies the usefulness of Dr. Hakala's work; from all appearances, the event study is more concerned simply with identifying abnormal market movement than in supporting the shareholders' causation allegations. Thus, we agree with the district court's negative assessment of Dr. Hakala's selection of event dates.

On appeal, the shareholders argue that the district court could only arrive at this conclusion by rejecting Dr. Hakala's testimony, and that by so doing it interposed itself as a fact-finder. It is true that a trial court should not “determine which of several competing scientific theories has the best provenance.” *Id.* at 85. If an expert has reached her conclusion “in a scientifically sound and methodologically reliable fashion,” *id.*, the differences “should be tested by the adversarial process,” *Milward*, 639 F.3d at 15. Moreover, the court should not rely on credibility determinations to resolve a disagreement between experts. See *Seahorse Marine Supplies, Inc. v. P.R. Sun Oil Co.*, 295 F.3d 68, 81 (1st Cir.2002) (“The ultimate  
92 \*92 credibility determination and the testimony's accorded weight are in the jury's province.”).

Here, Dr. Hakala stated on several occasions that he pre-selected relevant event dates without reference to the stock price, yet the district court specifically found, to the contrary, that Dr. Hakala had “cherry-picked unusually volatile days and made them the focus of the study.” *Bricklayers*, 853 F.Supp.2d at 188. The shareholders claim that the district court impermissibly discredited Dr. Hakala's testimony on this issue. This argument misses the point. The problem is not whether Dr. Hakala selected his event dates with reference to AOL's stock price. The problem is that the indisputably volatile dates that Dr. Hakala selected

were often unrelated to the shareholders' allegations, and therefore do not "help the trier of fact to understand the evidence or to determine a fact in issue." [Fed.R.Evid. 702\(a\)](#). The district court focused on this deficiency, and not on the mechanics of how Dr. Hakala selected these event dates. Consequently, we see no reason to address whether the district court made an impermissible credibility determination.

## 2. Overuse of Dummy Variables

We turn next to the district court's conclusion that Dr. Hakala overused dummy variables, which, according to the court, "artificially deflated the baseline volatility of AOL's stock price during the Class Period." [Bricklayers](#), 853 F.Supp.2d at 189. While our review of the record lends some support to the district court's assessment, there are countervailing factors suggesting that Dr. Hakala's exclusion of various dates during the Class Period affects only the weight, and not the admissibility, of his event study.

CSFB argued, and the district court agreed, that "Dr. Hakala's event study uses a much higher percentage of dummy variables than is considered acceptable in the financial econometric community." *Id.* at 188. We think, however, that in arriving at this conclusion, the court may have given insufficient weight to the shareholders' proffer. The shareholders offered scholarship, *see, e.g.,* Robert B. Thompson, II, et al., *The Influence of Estimation Period News Events on Standardized Market Model Prediction*, 63 *Acc. Rev.* 448, 466 (1988) ("[T]he distribution of security returns during periods in which *Wall Street Journal* news is released appears to differ systematically from the distribution of non-release period returns. This 'news-release' effect can be incorporated in models of the returns generating process by conditioning on news releases."), as well as expert testimony from Dr. M. Laurentius Marais<sup>10</sup> that supported Dr. Hakala's approach.

10 Dr. Marais submitted testimony rebutting Dr. Stulz's criticism of Dr. Hakala's use of dummy variables.

CSFB has identified articles that describe event study methodologies without mentioning the option of controlling for material news. *See, e.g.,* A. Craig MacKinlay, *Event Studies in Economics and Finance*, 35 *J. Econ. Literature* 13, 17–19 (describing various market models for event studies without mentioning a news-conditioned model, but noting that "[t]he use of other models is dictated by data availability"). The shareholders, meanwhile, point to academic event studies that do control for material news dates using a definition of "material news" narrower than Dr. Hakala's. *See, e.g.,* Richard Roll, *R*<sup>2</sup>, 43 *J. Fin.* 541, 558 (1988) (selecting \*93 material news from the Dow–Jones news service and *The Wall Street Journal*).

Ultimately, Dr. Hakala's approach may not be inconsistent with the methodology or goals of a regression analysis. A regression analysis seeks to isolate the effect that one variable has on another. Dr. Hakala's event study sought to isolate the effect of the general market conditions on AOL's stock price. He believes that "material news dates" have the potential to distort this relationship, and therefore excludes them from his analysis. Other market economists may disagree with the efficacy of this step or with the way that he defines materiality, but it is hard to see how it fails to follow the logic of regression studies. Indeed, CSFB's event study excludes certain dates for precisely the same reason. Nor do we consider the percentage of dummied-out dates dispositive of the issue. The district court was troubled by the fact that Dr. Hakala excluded 211 of the 388 dates in the study period. [Bricklayers](#), 853 F.Supp.2d at 188. That fact alone, however, does not negate the reliability of his study. The remaining 177 dates provided enough data to conduct a robust regression analysis. As Dr. Marais (the shareholders' expert on the issue of dummy



variables) noted, the important factor is not “the mechanistic and superficial percent of some universe of observations” that Dr. Hakala dummed out, but the “valid technical principles concerning the validity of the exercise.”

The district court noted two previous court opinions that disapproved of Dr. Hakala's use of dummy variables. We have held, however, that “the question of admissibility must be tied to the facts of a particular case.” *Milward*, 639 F.3d at 14–15 (citations omitted) (internal quotation marks omitted). The importance of that counsel is manifest here. Based on the record before us, Dr. Hakala's event studies in those two cases differed from this one in at least one key respect: in the other cases he dummed out dates on which “any news” about the company appeared. *Northfield Labs.*, 267 F.R.D. at 548; see also *Xcelera*, 2008 WL 7084626, at \*1. This is not a frivolous distinction, and the district court in *Xcelera* highlighted its importance: “Although the academic literature supports the use of dummy variables for events in which significant company-specific news is released, no peer-reviewed journal supports the view that dummy variables may be used on all dates on which any company news appears.” *Xcelera*, 2008 WL 7084626, at \*1. No one contends that Dr. Hakala dummed out every day in which AOL appeared in a news story, yet that was precisely the problem in *Northfield* and *Xcelera*.<sup>11</sup> Given that Dr. Hakala employed a different methodology for this case, *Northfield* and *Xcelera* are of limited value in assessing it.

<sup>11</sup> CSFB argues that Dr. Hakala employed the same “material news” standard in his event studies in *Northfield* and *Xcelera* as he did here. That may be true, but it does not address the fact that the courts in those cases specifically found that Dr. Hakala excluded any date containing company-specific news. No such finding exists in this case.

In *Bazemore v. Friday*, 478 U.S. 385, 400, 106 S.Ct. 3000, 92 L.Ed.2d 315 (1986) The Supreme Court observed that, “Normally, failure to include variables will affect the analysis' probativeness, not its admissibility.” Thus, while Dr. Hakala's use of dummy variables may, as defendants contend, have artificially deflated the baseline volatility of AOL's stock in his regression analysis, it may be a dispute that should be resolved by the jury.

CSFB launches one more assault on Dr. Hakala's use of dummy variables. It contends that his methodology fails under *Daubert* because it cannot be replicated. \*94 Dr. Hakala has performed three separate event studies related to the AOL merger, and each time he has dummed out more material news dates than before. Consequently, CSFB argues, his selection of material news dates is arbitrary and could not be replicated by another economist. We are not so sure.

*Daubert* suggests that a key question in determining whether a particular technique is scientific knowledge that will be useful to a jury is “whether it can be (and has been) tested.” *Daubert*, 509 U.S. at 593, 113 S.Ct. 2786. There, the Court was encouraging trial courts to limit expert testimony to falsifiable theories, meaning those “capable of empirical test.” *Id.* (quoting Carl Hempel, *Philosophy of Natural Science* 49 (1966)). Testing a particular theory will either reproduce consistent results, thus confirming the theory, or inconsistent results, thus casting doubt on it. In this case, Dr. Hakala has theorized that, given certain assumptions, AOL's stock experienced abnormal returns on fifty-seven event dates. One would test that theory by repeating his event study under the same conditions that he did. This would not be a difficult task, since Dr. Hakala has provided all of the necessary guidelines to recreate his event study.

Rather than put Dr. Hakala to the test, CSFB has simply argued that Dr. Hakala's techniques are unreproducible because of differences in the number of material news dates that he has

dummied out in successive event studies. But CSFB here is not comparing apples to apples. Only one of the three studies to which they refer is at issue here. One of the others was created in support of class action certification; the other was in connection with a different lawsuit. That fact alone could be enough to neuter CSFB's argument and leave such matters as fodder for cross-examination, not exclusion.

Ultimately, both the number of dates Dr. Hakala excluded from consideration and the methods he employed to select those dates create close questions. And while, as noted, appellant's arguments raise credible questions, we need not resolve this particular sub-issue because, as the district court concluded, the other three bases for excluding Dr. Hakala's testimony are sound.

### 3. Prior Disclosures

We have described an efficient market for the purpose of class action securities litigation as “one in which the market price of the stock *fully reflects all* publicly available information.” *In re PolyMedica Corp. Sec. Litig.*, 432 F.3d 1, 14 (1st Cir.2005). We have also explained that the relevant inquiry is whether the market is informationally efficient, *id.* at 16, meaning that “all publicly available information is impounded in [the] price” rapidly after it is disseminated. *Id.* at 14. The district court correctly applied this standard to Dr. Hakala's event study. Having established that AOL stocks traded in an efficient market in order to obtain class certification, the shareholders could not abandon that factual premise when proving loss causation. Yet several of the relevant events in Dr. Hakala's study are based on published references to information previously disclosed that, under an efficient market theory, would have already been incorporated into AOL's share price. The lag between the original disclosure and the event date ranged from one day to roughly a month. The majority of these disclosures occurred at least a

week before the event dates; thus, the event dates occurred long after an efficient market would have processed the news.

The shareholders respond that the event dates included new information that was not contained in the original disclosures. <sup>95</sup> We conclude, however, that while the disclosures made on the event dates did not merely parrot previously released information, they did no more than to provide gloss on public information, and thus permitted the district court to find that they would not have moved AOL's share price in an efficient market. *See In re Omnicom Grp.*, 597 F.3d at 512 (holding that the “negative characterization of already-public information” does not constitute a corrective disclosure of new information). For instance, Dr. Hakala included a February 2001 Lehman Brothers report on AOL. While this report downgraded its January 2001 buy or sell recommendation for AOL, it based this downgrade on information that was known the previous month. That Lehman Brothers reconsidered its initial appraisal of AOL's business, or lost confidence in AOL from one month to the next, does not demonstrate corrective information entering the market. *See id.* (“A negative journalistic characterization of previously disclosed facts does not constitute a corrective disclosure of anything but the journalists' opinions.”). The district court did not abuse its discretion in determining that this recurring problem affected the admissibility of Dr. Hakala's event study.

### 4. Confounding Factors

When proving loss causation in a securities fraud suit, plaintiffs “bear[ ] the burden of showing that [their] losses were attributable to the revelation of the fraud and not the myriad other factors that affect a company's stock price.” *In re Williams Sec. Litig.*, 558 F.3d 1130, 1137 (10th Cir.2009); *Dura*, 544 U.S. at 343, 125 S.Ct. 1627 (holding that a plaintiff does not show loss causation if the lower share price reflects “not the earlier misrepresentation, but changed economic

circumstances, changed investor expectations, new industry-specific or firm-specific facts, conditions, or other events, which taken separately or together account for some or all of that lower price”). Thus, when conducting an event study, an expert must address confounding information that entered the market on the event date.

This case deals with a highly publicized merger that captured the attention of the entire financial industry. There is no doubt that Dr. Hakala faced a “herculean task” in sorting through the continuous flow of information about AOL. *Bricklayers*, 853 F.Supp.2d at 190. We agree with the district court, however, that Dr. Hakala did not establish any reliable means of addressing this problem. Instead, he seemingly made a judgment call as to confounding information without any methodological underpinning.

In support of Dr. Hakala's treatment of confounding factors, the shareholders correctly point out that “even a statistical event study involves subjective elements.” *In re Xerox Corp. Sec. Litig.*, 746 F.Supp.2d 402, 412 (D.Conn.2010) (citations omitted) (internal quotation marks omitted). Nevertheless, a subjective analysis without any methodological constraints does not satisfy the requirements of *Daubert*. As the district court noted, “[i]t would be just as scientific to submit to the jurors evidence of defendants' alleged fraud and AOL's stock fluctuations and let them speculate whether the former caused the latter.” *Bricklayers*, 853 F.Supp.2d at 190; cf. *Milward*, 639 F.3d at 17–19 (admitting expert testimony based on a subjective “weight of the evidence” methodology, but identifying the established steps in this analysis and the factors used in analyzing the causal relationship). Dr. Hakala had tools at his disposal, such as intra-day trading analysis, to guide his analysis of confounding information.<sup>12</sup>

*Analysis* (2007), available at [http://www.nera.com/67\\_5197.htm](http://www.nera.com/67_5197.htm); Esther Bruegger & Frederick C. Dunbar, *Estimating Financial Fraud Damages with Response Coefficients*, 35 J. Corp. L. 11, 25 (2009) (“[C]ontent analysis' is now part of the tool kit for determining which among a number of simultaneous news events had effects on the stock price.”).

## 5. Bottom Line

Ultimately, we conclude that the district court did not abuse its discretion in excluding Dr. Hakala's testimony. While we may question its analysis with respect to dummy variables, the court's treatment of the remaining three issues is more than sufficient to satisfy our deferential review. *See Ruiz–Troche*, 161 F.3d at 83 (“[W]e will reverse a trial court's decision if we determine that the judge committed a meaningful error in judgment.” (citations omitted) (internal quotation marks omitted)).

Even conceding the aforementioned problems with Dr. Hakala's event study, however, the shareholders contend that the event study identified abnormal market movement, on certain key dates, that did not suffer from any methodological infirmities. Therefore, they claim, the district court abused its discretion by throwing out the good with the bad. True enough, some reviewing courts have found abuses of discretion where trial courts rejected mostly salvageable expert testimony for narrow flaws. *See City of Tuscaloosa v. Harcros Chems., Inc.*, 158 F.3d 548, 563 (11th Cir.1998) (reversing the exclusion of expert testimony in its entirety where only “a small portion of [the] data and testimony [was] fundamentally flawed”). Here, however, we confront the reverse situation—pervasive problems with Dr. Hakala's event study that, allegedly, still leave a few dates unaffected.

<sup>12</sup> Dr. Hakala could also have used content analysis. See, e.g., David Tabak, *Making Assessments About Materiality Less Subjective Through the Use of Content*

The district court was not obligated to prune away all of the problematic events in order to preserve Dr. Hakala's testimony. Out of fifty-seven event dates, the shareholders list five "key disclosures" that should survive the district court's order. The district court did not abuse its discretion in treating the entire event study as inadmissible given the overwhelming imbalance between unreliable and reliable dates. The burden of proof falls on the party introducing expert testimony. *Moore v. Ashland Chem. Inc.*, 151 F.3d 269, 276 (5th Cir.1998) ("The proponent need not prove to the judge that the expert's testimony is correct, but she must prove by a preponderance of the evidence that the testimony is reliable."). Requiring judges to sort through all inadmissible testimony in order to save the remaining portions, however small, would effectively shift the burden of proof and reward experts who fill their testimony with as much borderline material as possible. We decline to overturn the district court's ruling on this specious logic.

We also reject the shareholders' argument that CSFB ambushed them with new arguments at the *Daubert* hearing.<sup>13</sup> CSFB presented no new arguments at the *Daubert* hearing. Instead, it made a thorough presentation of the alleged problems of each event date. This should not have caught the shareholders off guard. The *Daubert* hearing occurred over three years after CSFB first challenged Dr. Hakala's expert testimony. During those three \*97 years, CSFB reiterated its arguments in expert reports, depositions, and briefings. It cited numerous examples of specific dates, but never claimed that those dates constituted the entirety of Dr. Hakala's flaws. The shareholders knew how CSFB would attack Dr. Hakala's event study, and they could have anticipated the scope of the attack. **B. Summary Judgment**

<sup>13</sup> The parties argue over which standard of review we should apply to this issue. We need not answer that question, as the

outcome is the same under any standard of review.

Our review of a grant of summary judgment is *de novo*, interpreting the record in the light most favorable to the nonmoving party. See *Henry v. United Bank*, 686 F.3d 50, 54 (1st Cir.2012). "Under [Federal Rule of Civil Procedure 56(a)], summary judgment is proper if the pleadings, depositions, answers to interrogatories, and admissions on file, together with the affidavits, if any, show that there is no genuine issue as to any material fact and that the moving party is entitled to a judgment as a matter of law." *Celotex Corp. v. Catrett*, 477 U.S. 317, 322, 106 S.Ct. 2548, 91 L.Ed.2d 265 (1986) (internal quotation marks omitted).

Although the district court awarded summary judgment to CSFB "even if [Dr. Hakala's event study] were admitted," *Bricklayers* at 191–92, we need not engage in such counter-factual analysis, see *Peguero–Moronta v. Santiago*, 464 F.3d 29, 34 (1st Cir.2006) ("We can affirm [the district court] on any basis available in the record...."). To sustain this suit, the shareholders needed to show a connection between CSFB's deceptive practices and the drop in AOL's stock price. The shareholders relied solely on Dr. Hakala's event study to satisfy this element. Without it, they cannot show a genuine dispute as to this issue. The district court did not need to tunnel into Dr. Hakala's event study for any evidence favorable to the shareholders' claim. The district court excluded Dr. Hakala's testimony in its entirety. We uphold that ruling. Thus, there is no evidence to sort through, and this complete lack of evidence compels a grant of summary judgment to CSFB.

### III. Conclusion

For the foregoing reasons, we *affirm* the district court's exclusion of the shareholders' expert testimony and consequently *affirm* its award of summary judgment to CSFB.

It is so ordered.

 casetext

# Exhibit 65

# Exhibit A



**IN THE UNITED STATES DISTRICT COURT  
FOR THE SOUTHERN DISTRICT OF TEXAS  
HOUSTON DIVISION****ENTERED**

September 27, 2023

Nathan Ochsner, Clerk

DELAWARE COUNTY EMPLOYEES	§	
RETIREMENT SYSTEM, individually and	§	
on behalf of all others similarly situated,	§	
	§	CIVIL ACTION NO. H-21-2045
Plaintiffs,	§	
v.	§	
	§	
CABOT OIL & GAS CORPORATION, et	§	
al.,		
Defendants.		

**MEMORANDUM AND OPINION**

This action arises under §§ 10(b) and 20(a) of the Securities Exchange Act of 1934, as amended by the Private Securities Litigation Reform Act of 1995, 15 U.S.C. §§ 78j(b), 78t(a), and its implementing regulation, Rule 10b–5, 17 C.F.R. § 240.10b–5. The named plaintiffs are two retirement plans that purchased common stock in Cabot Oil & Gas Corporation between February 22, 2016, and June 12, 2020: the Delaware County Employees Retirement System and the Iron Workers District Council (Philadelphia and Vicinity Retirement and Pension Plan) (“the Plans”). The Plans sued Cabot and three of its executive officers: Dan Dinges, the Chief Operating Officer; Scott Schroeder, the Chief Financial Officer; and Phil Stalnaker, the Vice President and Regional Manager (together, “the Cabot officers”). The Plans allege that Cabot and the Cabot officers publicly misrepresented Cabot’s compliance with the environmental laws in Susquehanna County, Pennsylvania, where its most significant well sites are located. According to the Plans, these misrepresentations kept Cabot’s stock price artificially inflated until the truth about Cabot’s noncompliance was revealed, causing the stock to plummet. The Plans allege millions of dollars in damages caused by Cabot’s misrepresentations and omissions.

The Plans move for certification of a class “consisting of all persons or entities who purchased or otherwise acquired Cabot . . . common stock between February 22, 2016 and June

Case 2:21-cv-00254 Document 1-1 Filed 01/27/23 Page 2 of 23  
12, 2020, inclusive . . . , and were damaged thereby.” (Docket Entry No. 134-1 at 7–8). The Plans also move to be appointed as class representatives and to have Robbins Geller Rudman & Dowd LLP and Kessler Topaz Meltzer & Check, LLP appointed as class counsel. (*Id.* at 8). Based on the parties’ pleadings, the motions and responses, the record, and the applicable law, the motions are granted. The reasons are set out below.

## **I. Background**

Cabot Oil & Gas Corporation is an oil and gas company publicly traded on the New York Stock Exchange. (Docket Entry No. 110 at ¶ 26). Cabot does hydraulic fracturing, or fracking, in the Marcellus Shale Deposit.<sup>1</sup> (*Id.* at ¶ 43). Most of Cabot’s production comes from the Marcellus Shale underneath Susquehanna County, Pennsylvania, and specifically the Dimock Township. (*Id.* at ¶¶ 30–31, 92).

Cabot began drilling in the Dimock Township in 2006. (*Id.* at ¶ 92). Sometime after, Dimock residents noticed sediment and “effervescence” in their drinking water. (*Id.* at ¶ 93). Some residents experienced strange symptoms, including tunnel vision, nausea and “bodily blotches.” (*Id.* at ¶ 96). One resident decided to hold a lighter to the water and “flames came flying out of the jug.” (*Id.* at ¶ 97). A residential water well near a Cabot drilling site spontaneously exploded. (*Id.* at ¶ 93). Dimock residents and the Pennsylvania Department of Environmental Protection (the “Pennsylvania Department”) suspected that Cabot was responsible. (*Id.* at ¶¶ 93–94).

The Pennsylvania Department investigated and concluded that methane gas was migrating from Cabot’s drilling sites into Dimock’s aquifer.<sup>2</sup> (*Id.* at ¶ 94). Cabot entered into a consent

---

<sup>1</sup> According to the complaint, “[f]racking is a multistep process that involves drilling down deep into the earth to shale rock deposits. Next, fluid is injected into the rock at high pressures to create fractures in the shale that release the natural gas otherwise trapped within the rock. The gas is then able to flow freely up to the surface to be captured by oil and gas operators . . . .” (Docket Entry No. 110 at ¶ 89).

<sup>2</sup> “Methane is the chief constituent of natural gas, which contains from 50% to 90% methane, and is a potentially explosive gas that burns readily in air.” (Docket Entry No. 110 at ¶ 90). Methane can escape a

Case 2:21-cv-00254 Document 47-1 Filed 01/27/23 Page 21 of 43  
order and agreement with the Pennsylvania Department, which was finalized in December 2010 (the “2010 Consent Order”). The 2010 Consent Order shut down Cabot’s operations in a nine-square mile area of Dimock called the “Dimock Box,” and required Cabot to remediate Dimock’s drinking water and Cabot’s wells in the Dimock Box to prevent further contamination. (*Id.*) Cabot also pledged under the 2010 Consent Order to comply with all applicable environmental laws and regulations. (*Id.* at ¶ 47).

The Plans allege that over the next decade, Cabot did not remediate the drinking water or its Dimock wells and continued to violate environmental laws. (*Id.* at ¶ 118). The Pennsylvania Department cited Cabot for 781 violations between January 2011 and March 2020, for wells both within the Dimock Box and elsewhere in Susquehanna County. (*Id.* at ¶ 142).

On February 11, 2020, a Pennsylvania grand jury recommended criminal charges against Cabot based on its findings that Cabot knowingly contaminated residential water supplies in Susquehanna County through its drilling and fracking. (*Id.* at ¶ 84). The grand jury’s findings were made public on June 15, 2020, when the Pennsylvania Attorney General released the “Presentment of Charges,” charging Cabot with fifteen criminal counts, including nine felonies. (*Id.* at ¶¶ 84, 120). The Attorney General charged Cabot under the Pennsylvania Clean Streams Law for knowingly discharging industrial waste, 35 P.S. § 691.301, and knowingly polluting Pennsylvania waters, 35 P.S. § 691.401. (*Id.* at ¶ 122). The charges were based not only on the Dimock wells covered by the 2010 Consent Order,<sup>3</sup> but also on other wells with similar pollution

---

well when there are problems with the cement that is poured into spaces between a wellbore (that is, a hole drilled in the ground) and its casing (that is, a pipe running through the hole). (*Id.* at ¶ 90). The cement is supposed to secure the pipe and fill gaps through which gasses can escape. (*Id.*) But when, for example, groundwater mixes with the cement and dilutes it, the cement can set imperfectly, allowing gasses like methane to escape. (*Id.* at ¶ 91). Cement problems were what caused Cabot’s wells to leak methane. (*Id.* at ¶ 110).

<sup>3</sup> These wells were: G Shields 1V, G Shields 2H, G Shields 4H, G Shields 5H, Costello 1V, Costello 2V, Gesford 4H, Gesford 8H, Ratzel 1H, Ratzel 2H, Ratzel 3V, Ely 4H, and Ely 6H. (Docket Entry No. 110 at ¶ 121). The Pennsylvania Department issued notices of violation for the G Shields wells on October 20,

Case 2:21-cv-00025 Document 1-1 Filed 01/27/23 Page 45 of 143  
problems that Cabot had drilled later, outside Dimock.<sup>4</sup> (*Id.* at ¶ 120; Docket Entry No. 142-15 at 3–7, 20). The charges were based on Cabot’s operation of its wells “through at least June 11, 2018.”<sup>5</sup> (Docket Entry 142-15 at 3–7, 20). The Attorney General also charged Cabot with knowingly failing to comply with the 2010 Consent Order and other orders of the Pennsylvania Department. (Docket Entry No. 110 at ¶ 123). This charge was based on Cabot’s conduct “on or about December 15, 2010, through January 9, 2020.” (*Id.*). Cabot’s stock price dropped by over 3% after the Attorney General released the Presentment of Charges. (*Id.* at ¶ 263).

Cabot issued several public statements during the class period (February 22, 2016, to June 12, 2020). The parties have organized these statements into three categories:<sup>6</sup>

### **Category 1 Statements.**

Category 1 covers statements that Cabot made in its Form 10-Qs and Form 10-Ks in 2015 and 2016.<sup>7</sup> (Docket Entry No. 110 at ¶ 187–88; Docket Entry No. 134-1 at 12). The relevant representations are that Cabot: had received a notice of violation from the Pennsylvania Department in September 2011 for failing to prevent the migration of methane gas into

---

2011, for the Costello and Gesford wells on June 16, 2014, for the Ratzel wells on December 19, 2014, and for the Ely wells on March 21, 2018. (*Id.* at ¶ 126).

<sup>4</sup> The wells outside Dimock are the Howell well pad in Auburn Township, the Jeffers Farm wells in Harford Township, and the Powers M wells in Auburn Township. (Docket Entry No. 142-15 at 20). The Pennsylvania Department issued notices of violation for the Howell wells on March 16, 2017 and June 20, 2017, (*id.*), for the Jeffers Farm wells on November 16, 2017, (*id.* at 21), and for a Powers M well on October 18, 2019, (*id.*).

<sup>5</sup> June 11, 2018, is the date when the Pennsylvania Department sent a private letter to Cabot detailing its continuing violations of the 2010 Consent Order. The Pennsylvania Department defines “continuing violation” as a “violation that was noted in a previous inspection that is still continuing through current inspection.” (Docket Entry No. 110 at ¶ 146). In the letter, the Pennsylvania Department denied Cabot’s request to drill new wells within the Dimock Box because “Cabot is not in compliance with its obligation under the 2010 [Consent Order].” (Docket Entry No. 110 at ¶ 138).

<sup>6</sup> The court pruned the universe of alleged misrepresentations in its order on Cabot’s motion to dismiss, eliminating claims based on “nonactionable opinions”—namely, that Cabot “believe[s] that it substantially compl[ies] with the Clean Water Act and related federal and state regulations.” (Docket Entry No. 118 at 7, 24).

<sup>7</sup> A Form 10-Q is a quarterly financial report, and a Form 10-K is an annual financial report, required by the Securities and Exchange Commission pursuant to section 13 or 15(d) of the Securities Exchange Act of 1934. 15 U.S.C. §§ 78m, 78(d).

Case 2:21-cv-00254 Document 1-1 Filed 01/27/23 Page 5 of 23  
Susquehanna County groundwater; was “engaged with the [Pennsylvania Department] in investigating the incident and ha[s] performed appropriate remediation efforts, including the provision of alternative sources of drinking water to affected residents”; believed “the source of methane has been remediated and [that it was] working with the [Pennsylvania Department] to reach agreement on the disposition of this matter”; had received a proposed consent order that, if finalized, would result in a civil penalty of between \$100,000 and \$300,000; and would continue to work to bring the matter to a close. (Docket Entry No. 110 at ¶¶ 187–88).

The September 2011 notice of violation and proposed consent order arose from Cabot’s operations at the Stalter well site in Lenox Township, which is within Susquehanna County. (*Id.* at ¶¶ 149, 187; Docket Entry No. 142-5).

### **Category 2 Statements.**

Category 2 comprises statements that Cabot made in its 2016 Form 10-K, filed on February 27, 2017. The statements provide an update on the September 2011 notice of violation and proposed consent order that were the subject of the Category 1 Statements:

On November 12, 2015, we received a proposed Consent Order and Agreement from the Pennsylvania Department of Environmental Protection (PaDEP) relating to gas migration allegations in an area surrounding several wells owned and operated by us in Susquehanna County, Pennsylvania. The allegations relating to these wells were initially raised by residents in the area in August 2011. We received a Notice of Violation from the PaDEP in September 2011 for failure to prevent the migration of gas into fresh groundwater sources in the area surrounding these wells. Since then, we have been engaged with the PaDEP in investigating the incident and have performed appropriate remediation efforts, including the provision of alternative sources of drinking water to affected residents. We believe the source of methane has been remediated and we entered into a Consent Order and Agreement with the PaDEP on December 30, 2016. We agreed to pay a civil monetary penalty in the amount of approximately \$0.3 million and to continue to provide alternative sources of drinking water to affected residents until the affected water supplies are permanently restored. Further, the related gas well is being permanently plugged. Following the plugging of the gas well, additional monitoring will be required to ensure the source of methane has been remediated. Cabot continues to work with the PaDEP to bring this matter to a close.

(Docket Entry No. 110 at ¶ 188).

### **Category 3 Statements.**

There are two separate substantive components to the Category 3 Statements. The first is a press release that Cabot published at 6:38 a.m. on July 26, 2019. (Docket Entry No. 142-1 at ¶ 32). The second is the disclosure of notices of violation that Cabot had received in June 2017 and November 2017 relating to the Howell and Jeffers Farm wells that Cabot had drilled in Susquehanna County in the Auburn and Harford Townships, respectively. (Docket Entry No. 110 at ¶¶ 127, 129). Cabot first disclosed these notices of violation in its 2Q19 Form 10-Q, filed on July 26, 2019, at 11:42 a.m. (Docket Entry No. 142-1 at ¶ 31).

The relevant part of the press release (the “Guidance Update”) follows.

Cabot has provided its third quarter 2019 production guidance range of 2,360 to 2,410 Mmcfe per day. The Company has also adjusted its 2019 production growth guidance to a range of 16 to 18 percent (24 to 26 percent on a debt-adjusted per share basis) due in large part to a change in the operating plan resulting from a unique opportunity to acquire acreage adjacent to an eight-well pad, allowing the Company to increase the total lateral footage on the pad by approximately 28,000 feet (increasing the average lateral length per well from 8,950 feet to 12,450 feet). “This increase in lateral lengths will improve the capital efficiency and economics of the pad; however, the longer cycle time will result in a delay in the wells being placed on production, pushing out the production contribution from this pad to late December or early January,” said Dinges. Cabot has updated its 2019 capital budget to a range of \$800 million to \$820 million to reflect the incremental drilling and completion activity on the previously referenced eight-well pad and an increase in drilling activity for the year by four net wells resulting from continued efficiency gains on the Company’s three fully contracted drilling rigs.

...

Cabot has provided its preliminary 2020 production growth guidance of five percent (seven to eight percent on a debt-adjusted per share basis). This production growth is based on a preliminary capital budget range of \$700 million to \$725 million. The Company’s 2020 program is expected to deliver \$375 million to \$400 million of free cash flow at a \$2.50 NYMEX price and \$525 million to \$550 million of free cash flow at a \$2.75 NYMEX price.

(Docket Entry No. 142-14 at 7–8).

The relevant part of the Form 10-Q is:

On June 17, 2019, we received two proposed Consent Order and Agreements (“CO&A”) from the Pennsylvania Department of Environmental Protection (PaDEP) relating to gas migration allegations in areas surrounding several wells owned and operated by us in Susquehanna County, Pennsylvania. The allegations relating to these wells were initially raised by residents in the area in March and

June 2017, respectively, in the form of complaints about their drinking water supply. Since then, we have been engaged with the PaDEP in investigating the incidents and have performed appropriate remediation efforts, including the provision of alternative sources of drinking water to the affected residents. We received Notices of Violation (“NOV”) from the PaDEP in June and November 2017, respectively, for failure to prevent the migration of gas into fresh groundwater sources in the area surrounding these wells. With regard to the June 2017 NOV, we believe these water quality complaints have been resolved, and we are working with the PaDEP to reach agreement on the disposition of this matter. The proposed CO&A is the culmination of this effort and, if finalized, would result in the payment of a civil monetary penalty in an amount likely to exceed \$100,000, up to approximately \$215,000. We will continue to work with the PaDEP to finalize the CO&A, and to bring this matter to a close. With regard to the November 2017 NOV, the proposed CO&A, if finalized as drafted, would require Cabot to submit a detailed written remediation plan, continue water sampling and other investigative measures and restore or replace affected water supplies and would result in the payment of a civil monetary penalty in an amount likely to exceed \$100,000, up to approximately \$355,000. We will continue to work with the PaDEP to finalize the CO&A, and to complete the ongoing investigation and remediation.

(Docket Entry No. 151-27 at 30).

Cabot’s stock price declined on July 26, 2019, the day of the Category 3 Statements, bottoming out 12% lower than when the market opened.<sup>8</sup> (Docket Entry No. 110 at ¶ 260).

## II. The Legal Standards

### A. Class Certification

Under Rule 23(a), plaintiffs seeking class certification must satisfy four elements: (1) numerosity, (2) commonality, (3) typicality, and (4) adequacy of representation. Fed. R. Civ. P. 23(a); *Wal-Mart Stores, Inc. v. Dukes*, 564 U.S. 338, 349 (2011).

Numerosity means that “the class is so numerous that joinder of all members is impracticable.” Fed. R. Civ. P. 23(a)(1). Commonality means that “there are questions of law or fact common to the class.” *Id.* 23(a)(2). Typicality means that “the claims or defenses of the

---

<sup>8</sup> The Plans use the Category 3 Statements not just as alleged misrepresentations but also as alleged “corrective disclosures,” that is, revelations that showed the public that the Category 1 and Category 2 statements were false. As explained below, “corrective disclosures” are often important in securities class actions to demonstrate that a company’s misrepresentations kept its stock price artificially high. If the stock price falls after the corrective disclosure, it can often be inferred that the misrepresentations had a “price impact.” See *Pub. Emps. Ret. Sys. of Miss. v. Amedisys, Inc.*, 769 F.3d 313, 320–21 (5th Cir. 2014); *Goldman Sachs Group, Inc. v. Ark. Teacher Retirement Sys.*, 594 U.S. ----, 141 S. Ct. 1951, 1961 (2021).



Case 2:21-cv-00055 Document 47-1 Filed 01/27/23 Page 23 of 43  
representative parties are typical of the claims or defenses of the class.” *Id.* 23(a)(3). Adequacy means that the representative party and the named class counsel “will fairly and adequately protect the interests of the class.” *Id.* 23(a)(4).

Once the plaintiffs satisfy those four elements, they must further show that the class action falls within at least one of the following three categories under Rule 23(b): (1) cases in which prosecuting separate actions by or against individual class members would create a risk of inconsistent adjudication; (2) cases in which “the party opposing the class has acted or refused to act on grounds that apply generally to the class,” so that final injunctive or declaratory relief is appropriate with respect to the class as a whole; or (3) cases in which “questions of law or fact common to class members predominate over any questions affecting only individual members” and the “class action is superior to other available methods for fairly and efficiently adjudicating the controversy.” *Id.* 23(b).

The Rule 23 analysis is “rigorous,” and “a district court must detail with sufficient specificity how the plaintiff has met the requirements of Rule 23.” *Chavez v. Plan Benefit Servs., Inc.*, 957 F.3d 542, 545 (5th Cir. 2020) (quoting reference omitted). “‘Rule 23 does not set forth a mere pleading standard.’” *Id.* (quoting *Dukes*, 564 U.S. at 350). “A party seeking class certification must affirmatively demonstrate his compliance with the Rule—that is, he must be prepared to prove that there are *in fact* sufficiently numerous parties, common questions of law or fact,’ and so on.” *Id.* (quoting *Dukes*, 564 U.S. at 350). A district court must often “probe behind the pleadings” and reach considerations “‘that are enmeshed in the factual and legal issues’ of the case.” *Id.* (first quoting *Gen. Tel. Co. of the Sw. v. Falcon*, 457 U.S. 147, 160 (1982), then quoting *Dukes*, 564 U.S. at 351). The court must “understand the claims, defenses, relevant facts, and applicable substantive law in order to make a meaningful determination.” *Flecha v. Medcredit, Inc.*, 946 F.3d 762, 766 (5th Cir. 2020).

## B. The Securities and Exchange Act

Section 10(b) of the Securities and Exchange Act of 1934 makes it “unlawful for any person, directly or indirectly, . . . [t]o use or employ, in connection with the purchase or sale of any security registered on a national securities exchange . . . any manipulative or deceptive device or contrivance in contravention of such rules and regulations as the [Securities and Exchange Commission] may prescribe as necessary or appropriate in the public interest or for the protection of investors.” 15 U.S.C. § 78j(b). Rule 10b–5 implements § 10(b) by prohibiting, among other things, the making of any “untrue statement of material fact” or the omission of any material fact “necessary in order to make the statements . . . not misleading.” 17 C.F.R. § 240.10b–5(b).

A plaintiff may recover damages under § 10(b) by showing: (1) a material misrepresentation or omission by the defendant; (2) scienter; (3) a connection between the misrepresentation or omission and the purchase or sale of a security; (4) reliance upon the misrepresentation or omission; (5) economic loss; and (6) loss causation. *R2 Invs. LDC v. Phillips*, 401 F.3d 638, 641 (5th Cir. 2005).

As in many securities class actions, the issue here is the intersection between the reliance element of a § 10(b) claim and the predominance requirement of Rule 23(b)(3). An individual plaintiff may establish reliance by showing “that he was aware of a defendant’s misrepresentation and engaged in a transaction based on that misrepresentation.” *Goldman Sachs Group Inc. v. Ark. Teacher Ret. Sys.*, 594 U.S. ----, 141 S. Ct. 1951, 1958 (2021). But this individualized inquiry is impracticable in a class action. To prove that individual questions of reliance do not predominate over questions common to the class, class-action plaintiffs invoke the presumption established in *Basic Inc. v. Levinson*, 485 U.S. 224, 248 (1988). *Id.* at 1959. Under the *Basic* presumption, courts presume that stock trading in an efficient market incorporates into its price all public, material information—including material misrepresentations—and that investors rely on the

Case 2:21-cv-00855 Document 1-2 Filed 09/29/23 in TXSD Page 12 of 24  
integrity of the market price when they choose to buy or sell that stock. See *Erica P. John Fund, Inc. v. Halliburton Co.* (“*Halliburton I*”), 563 U.S. 804, 813 (2011).

To establish the *Basic* presumption, plaintiffs must prove: (1) that the alleged misrepresentation was publicly known; (2) that it was material; (3) that the stock traded in an efficient market; and (4) that the plaintiff traded the stock between the time the misrepresentation was made and when the truth was revealed. *Halliburton Co. v. Erica P. John Fund, Inc.* (“*Halliburton II*”), 573 U.S. 258, 268 (2014). Once the presumption is established, it can be rebutted if the defendant proves, by a preponderance of the evidence, *Goldman*, 141 S. Ct. at 1960, “that an alleged misrepresentation did not actually affect the market price of the stock,” *Halliburton II*, 573 U.S. at 284. This can be done through “[a]ny showing that severs the link between the alleged misrepresentation and either the price received (or paid) by the plaintiff, or his decision to trade at a fair market price.” *Basic*, 485 U.S. at 248. Courts must consider “all probative evidence” of price impact, “aided by a good dose of common sense.” *Goldman*, 141 S. Ct. at 1960 (quoting reference omitted). “The district court’s task is simply to assess all the evidence of price impact—direct and indirect—and determine whether it is more likely that not that the alleged misrepresentations had a price impact.” *Id.*

The class-certification question of whether the *Basic* presumption has been rebutted by evidence of no price impact overlaps with the merits questions of materiality, reliance, and loss causation. See *id.* at 1961. Nonetheless, district courts must consider all probative evidence. *Id.* In doing so, they must “resist[] the temptation” to draw merits conclusions. *Id.* at 1961, n.2 (quoting reference omitted).

When a plaintiff’s theory is that the defendant’s misrepresentations or omissions kept its stock artificially inflated, price impact may be shown on the back end. See *id.* at 1961. Front-end price impact may be inferred from a back-end price drop when a corrective disclosure shows that the defendant’s previous statements were untrue or that the defendant failed to disclose the truth,

Case 2:21-cv-00855 Document 1-2 Filed 09/29/23 in TXSD Page 12 of 24  
and when the stock price falls after the truth is revealed. *See id.* However, a back-end price drop supports this inference only when the corrective disclosure “matches” the earlier misrepresentations or omissions. *See id.* If there is a “mismatch” between the contents of the corrective disclosure and the misrepresentations or omissions, the inference of price impact is weaker. *See id.* The Supreme Court has given the example of “when the earlier misrepresentation is generic (e.g., ‘we have faith in our business model’) and the later corrective disclosure is specific (e.g., ‘our fourth quarter earnings did not meet expectations’).” *Id.* That said, a corrective disclosure need not “precisely mirror” the misrepresentation. *Alaska Elec. Pension Fund v. Flowserve Corp.*, 572 F.3d 221, 230 (5th Cir. 2009). The two need only be “related” or “relevant” to one another. *Pub. Employees Ret. Sys. of Mississippi, Puerto Rico Teachers Ret. Sys. v. Amedisys, Inc.*, 769 F.3d 313, 321 (5th Cir. 2014).

### III. Analysis

Cabot does not contest that the Plans have satisfied Rule 23(a). The putative class is numerous. Cabot stock traded on the New York Stock Exchange and had an average weekly trading volume of 33.49 million shares during the class period. (Docket Entry No. 134-1 at 14). *See In re Enron Corp. Sec.*, 529 F. Supp. 2d 644, 672 (S.D. Tex. 2006) (numerosity is “generally assumed” in a class action involving nationally traded securities); *Mullen v. Treasure Chest Casino, LLC*, 186 F.3d 620, 624 (5th Cir. 1999) (“100 to 150 members[] is within the range that generally satisfies the numerosity requirement.”). Whether Cabot’s representations were materially false and impacted the stock price are common questions of law and fact. (Docket Entry No. 134-1 at 15). *See Dukes*, 564 U.S. at 359 (“[E]ven a single common question will do.”) (internal quotation marks omitted) (alteration adopted). The Plans’ claims that Cabot and its officers made misrepresentations that caused them damages are typical of the class members’ claims. (Docket Entry No. 134-1 at 16). *See In re Dynegy, Inc. Sec. Litig.*, 226 F.R.D. 263, 287–88 (S.D. Tex. 2005). The Plans and their counsel are adequate. The Plans’ interests are aligned

Case 2:21-cv-00855 Document 1-2 Filed 09/29/23 In: TXSD Page 13 of 243  
with those of the class and there are no conflicts between the Plans and the class members. (Docket Entry No. 134-1 at 17). *See In re Taxable Mun. Bond Sec. Litig.*, 51 F.3d 518, 522 (5th Cir. 1995); *Buettgen v. Harless*, 2011 WL 1938130, at \*5 (N.D. Tex. May 19, 2011). Finally, the Plans' counsel is competent and experienced. (Docket Entry Nos. 134-5, 134-6).

Cabot argues only that the Plans have not satisfied Rule 23(b)(3) because individualized reliance questions predominate. Cabot does not argue that the *Basic* presumption is absent, but rather argues that the presumption is rebutted by evidence that its representations about environmental compliance had no impact on its stock price. Cabot's many arguments can be roughly organized into two buckets. First, Cabot argues that the price drop following the Presentment of Charges is not statistically significant. Second, Cabot argues that there is a "mismatch" between its representations about environmental compliance and the two corrective disclosures: (1) the July 26, 2019, press release and Form 10-Q; and (2) the Presentment of Charges. Each argument is analyzed below.

#### **A. Statistical Significance**

Cabot's stock price fell 3% the day the Pennsylvania Attorney General released the Presentment of Charges. (Docket Entry No. 110 at ¶ 263). The parties' experts disagree whether this drop was statistically significant. Cabot's arguments assume that if the price drop was statistically insignificant, then that equals no price impact. The parties dispute the validity of that assumption as well. The court addresses the validity of Cabot's assumption before analyzing the evidence on both sides of the question of statistical insignificance. But first, the dual role of the Category 3 Statements deserves clarification.

The Category 3 Statements are allegedly both misrepresentations and corrective disclosures. If Cabot is correct that its stock price did not materially move following the Presentment of Charges, then there would be no evidence that the Category 3 Statements had a price impact. The Statements would then be relegated to the single duty of corrective disclosures,

Case 2:21-cv-00855 Document 12-1 Filed 09/29/23 in TXSD Page 13 of 243  
but that does not determine whether the Rule 23(b)(3) predominance requirement is unmet. To make that showing, Cabot would have to also prove that the Category 1 and 2 Statements had no price impact. As evidence that those statements had price impact, the Plans rely not only on the price drop following the Presentment of Charges, but also on the drop following the Category 3 Statements. Cabot's attacks on that evidence are addressed in the "mismatch" section, Part III.B, *infra*.

Cabot's underlying assumption is that the conclusion of no price impact follows from the statistical insignificance of a back-end price drop. The Plans say no, that as a matter of logic, statistical insignificance does not preclude price impact. (Docket Entry No. 151 at 20–21). They rely on *Rooney v. EZCORP, Inc.*, 330 F.R.D. 439, 450 (W.D. Tex. 2019), and *Monroe County Employees' Retirement System v. Southern Company ("Southern")*, 332 F.R.D. 370 (N.D. Ga. 2019). In *Rooney*, the court rejected the defendants' argument that the lack of a statistically significant price adjustment following a corrective disclosure established that those prior misrepresentations had no price impact. *Rooney*, 330 F.R.D. at 450. The court explained that "that is not how hypothesis testing works. A statically significant price adjustment following a corrective disclosure is evidence the original misrepresentation did, in fact, affect the stock price. The converse, however, is not true—the absence of a statistically significant price adjustment does *not* show the stock price was unaffected by the misrepresentation." *Id.* The *Southern* court similarly noted that, "[i]n recognition of [a] basic truism of statistics, courts routinely reject the argument that a non-statistically significant stock price decline proves an absence of price impact." 332 F.R.D. at 394. The court held that "[a] non-statistically significant decline simply does not 'sever the link' between the alleged misrepresentations and corrective disclosures." *Id.* at 395 (quoting *Basic*, 485 U.S. at 248).

The Plans' expert, Dr. Steven P. Feinstein, offers this explanation:

Price impact and price movement are not the same thing. Information that prevents a stock price decline has price impact even though there may be no movement in

the stock price. Therefore, while a statistically significant positive price reaction in response to misrepresentations may demonstrate price impact, a nonsignificant reaction, or even no price movement at all, does not prove there was no price impact.

(Docket Entry No. 151-2 at ¶ 25). When Cabot’s expert, Lucy P. Allen, was asked in her deposition whether she agreed with this reasoning, she testified that the statistical insignificance of a price drop is “evidence that there is no price impact.” (Docket Entry No. 151-3 at 85).

The reasoning of the *Rooney* and *Southern* courts is persuasive, and the parties have not pointed to convincing contrary authority. If Cabot is right that the price drop following the Presentment of Charges was statistically insignificant, that would be evidence that the misrepresentations had no price impact, but it would not be dispositive.

Here, the question of statistical significance requires the court to decide which sides’ expert is more likely correct. Before *Goldman*, district courts often shied away from engaging in a “battle of the experts” at the class-certification stage. *See, e.g., Southern*, 332 F.R.D. at 387–88 (“[T]he Court declines to ‘engage in the parties’ battle of the experts’ with respect to *Cammer* factor five . . . .”); *Zwick Partners, LP v. Quorum Health Corp.*, 2019 WL 1450546, at \*14 (M.D. Tenn. Mar. 29, 2019) (dueling expert reports “create[] a factual dispute . . . which involves complicated questions of causation better left until trial or at the earliest a summary judgment proceeding”). But the Supreme Court recognized that “most securities-fraud class actions” involve “competing expert evidence on price impact.” *Goldman*, 141 S. Ct. at 1963. It instructed district courts “to assess all the evidence of price impact—direct and indirect—and determine whether it is more likely than not that the alleged misrepresentations had a price impact.” *Id.* This task requires courts to keep in mind the importance of “common sense” in evaluating expert testimony on price impact. *Id.* at 1960; *see also In re Allstate Corp. Sec. Litig.*, 966 F.3d 595, 613 n.6 (7th Cir. 2020) (the price-impact analysis should be “aided by a good dose of common sense”) (quoting reference omitted).



The experts' disagreement about statistical significance appears to stem from a methodological difference. Dr. Feinstein applied the "Newey-West Methodology" to account for COVID-19-related market volatility, which he calls "heteroskedasticity."<sup>9</sup> (Docket Entry No. 134-3 at ¶¶ 137–140). He attributes Allen's conclusion of statistical insignificance to her failure to account for this market volatility, in addition to other alleged analytical flaws.<sup>10</sup> (Docket Entry No. 151-2 at ¶ 98). Allen responds that the Newey-West methodology is "non-standard" and "yields clearly erroneous and unreasonable results." (Docket Entry No. 142-1 at ¶ 66). She asserts that Dr. Feinstein's event study<sup>11</sup> "yields the unreasonably [*sic*] result that small excess returns were statistically significant while much larger excess returns during the same time period were not statistically significant." (*Id.* at ¶ 69).

Dr. Feinstein defends his methodology, asserting that the Newey-West methodology "is a peer-reviewed, published, and generally accepted methodology to correct for heteroskedasticity," and that Allen's own firm has endorsed the methodology. (Docket Entry No. 151-2 at ¶¶ 30–36, 108). He explains that the results of his methodology are not "erroneous or unreasonable" simply because "the biggest returns . . . are not always the most significant." (*Id.* at ¶ 107). "This is,"

---

<sup>9</sup> Investopedia explains that "heteroskedasticity" is "when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related to prior time periods, are non-constant." Investopedia, *Heteroscedasticity Definition: Simple Meaning and Types Explained*, <https://www.investopedia.com/terms/h/heteroskedasticity.asp>.

<sup>10</sup> The other alleged flaws include: "failing to correctly remove the effects of Cabot from her market and peer-company indices, testing the wrong dates, omitting dates without explanation, using inconsistent estimation periods, and using an erroneous return that 'dramatically biases Ms. Allen's results toward findings of event nonsignificance' . . . ." (Docket Entry No. 151 at 20 (citing Docket Entry No. 151-3 at ¶¶ 13–14, 37–48)).

<sup>11</sup> The court in *Erica P. John Fund, Inc. v. Halliburton Co.*, 309 F.R.D. 251, 262 (N.D. Tex. 2015), explained that "[a]n event study is generally comprised of two parts: (1) a calculation of the market-adjusted price change in the issuer's share price at the time the corrective disclosure became public . . . ; and (2) a determination of whether the corrective disclosure is among the . . . news that affected the price on the date the disclosure became public . . . by comparing the magnitude of the market-adjusted change in [Halliburton's] share price on the date of the corrective disclosure with the historical record of the daily, market-adjusted ups and downs in [Halliburton's] share price."

Case 2:21-cv-00855 Document 12-1 Filed 09/29/23 in TXSD Page 16 of 243  
instead, “a common and expected result when correcting for changing volatility, i.e., heteroskedasticity.” (*Id.* at ¶ 101).

Comparing Dr. Feinstein’s analysis with Ms. Allen’s leads to the conclusion that Dr. Feinstein’s is more reliable because it accounts for COVID-19-related market volatility. Although Allen acknowledged in her deposition that heteroskedasticity can cause errors in a regression analysis, she admits she did not test for heteroskedasticity in her analysis in this case. (Docket Entry No. 151-3 at 108, 111, 114). Ms. Allen’s criticism of Dr. Feinstein’s use of the Newey-West methodology is undermined by her own firm’s endorsement of the methodology in regression models. (*See* Docket Entry No. 151-2 at ¶¶ 99–100). Although Ms. Allen contends that her firm has not endorsed the Newey-West methodology “in the context” that Dr. Feinstein used it here, (Docket Entry No. 151-3 at 117), she does not explain why that is inappropriate. Instead, Ms. Allen criticizes Dr. Feinstein’s outputs and notes that he did not apply the Newey-West methodology in an event study he did for other, unrelated litigation, (*see* Docket Entry No 142-1 at ¶¶ 67–78). The court credits Dr. Feinstein’s conclusion that the residual return<sup>12</sup> following the Presentment of Charges is statistically significant. (Docket Entry No. 151-2 at ¶ 68).

Even if Cabot had shown that the price drop following the Presentment of Charges was statistically insignificant, common-sense considerations would discount the probative value of that showing. The announcement of felony criminal charges against Cabot based on its operations in Susquehanna County—which was by far its most productive region—would likely impact its stock price. The evidence shows that Cabot’s noncompliance with environmental laws and the 2010 Consent Order in Susquehanna County significantly affected its financials. (*See* Docket Entry No. 151 at 14–17). Cabot could not operate its existing wells or drill new wells in the Dimock Box until it satisfied its obligations under the 2010 Consent Order. (Docket Entry No. 110 at ¶ 138).

---

<sup>12</sup> Dr. Feinstein defines “residual return” as “the stock return after removing both the market and industry effects.” (Docket Entry No. 151-2 at ¶ 92).

The record also includes several examples of market commentary linking the June 15, 2020, price drop to the charges against Cabot. (See Docket Entry No. 151 at 23–24). See *Ark. Teacher Ret. Sys. v. Goldman Sachs Group, Inc.*, 77 F.4th 74, 104 (2d Cir. 2023) (“[M]arket commentary can provide insight into the kind of information investors would rely upon in making investment decisions—and therefore can serve as indirect evidence of price impact. . . .”).

Ms. Allen emphasizes that market analysts—as opposed to news outlets—generally did not report on the charges. But she admits that JP Morgan analysts did report on the charges two days after they were announced, and that the analysts changed their price targets for Cabot shortly thereafter. (Docket Entry No. 142-1 at ¶¶ 55–56). Dr. Feinstein points out that Ms. Allen overlooked a second analyst’s prediction on the day the Presentment of Charges was announced that the charges would have a “slightly negative” impact on Cabot’s stock price. (Docket Entry No. 151-2 at ¶ 71). In light of this evidence, Cabot’s argument that “analysts did not view the criminal charges as information material to Cabot’s stock price” is unpersuasive. (Docket Entry No. 141-1 at ¶ 54).

Cabot’s argument that the Presentment of Charges was immaterial because the potential fines were small compared to Cabot’s revenues is also unpersuasive. (Docket Entry No. 142 at 25–26). The Plans correctly point out that the potential fines were only one consequence of the company’s revelations. Other more serious consequences of “the revelation of Cabot’s longstanding (but undisclosed) failure to remediate known violations and gas migration incidents” included “fines, regulatory scrutiny, criminal charges, reputational damages, and reduced gas production.” (Docket Entry No. 151 at 25).

The price drop following the Presentment of Charges was statistically significant.

## **B. Mismatch**

Cabot next argues that there is a “mismatch” between its representations about environmental compliance and the two corrective disclosures. The Plans allege that the Category

3 Statements are corrective of the Category 1 and 2 Statements, and that the Presentment of Charges are corrective of the Category 1, Category 2, and Category 3 Statements. The court first analyzes whether there is a mismatch between the Category 3 Statements, as corrective disclosures, and the Category 1 and 2 Statements, then analyzes whether there is a mismatch between the Presentment of Charges and the Category 1, 2, and 3 Statements. In doing so, the court keeps in mind that corrective disclosures need not “precisely mirror” misrepresentations. *Alaska Elec. Pension Fund*, 572 F.3d at 230. The two need only be “related to” or “relevant to” one another. *Amedisys*, 769 F.3d at 321.

### 1. The Category 3 Statements as Corrective Disclosures

Cabot argues that there is a mismatch between the Category 3 Statements and the Category 1 and 2 Statements because the notices of violation disclosed in the 2Q10 Form 10-Q related to the Howell Wells and Jeffers Farms Pad 2 Wells, while the Category 1 and 2 Statements related to the Stalter Wells. (Docket Entry No. 142 at 14–15). Cabot further argues that the Guidance Update issued the same day as the Form 10-Q “had nothing to do with” its environmental compliance at any of the wells. (*Id.*). Cabot’s argument about the Form 10-Q is unpersuasive, but the court agrees that there is a mismatch between the Guidance Update and the Category 1 and 2 Statements.

There is no mismatch between the 2Q10 Form 10-Q and the Category 1 and 2 Statements. It would have been apparent to a careful observer that the violations disclosed in the 2Q10 Form 10-Q were distinct from the violation that was the subject of the Category 1 and 2 Statements. The 2Q10 Form 10-Q disclosed notices of violation received in 2017 based on complaints from residents in 2017. (Docket Entry No. 151-27 at 30). By contrast, the Form 10-Ks addressed Statements in Categories 1 and 2 that disclosed a notice of violation received in 2011 based on complaints from residents in 2011. (Docket Entry No. 110 at ¶¶ 187–88). However, it is not facially apparent from these documents that the distinct violations concerned different well sites.

Both the 2Q10 Form 10-Q and the Form 10-Ks in Categories 1 and 2 refer generally to “several wells owned and operated by us in Susquehanna County, Pennsylvania. (Docket Entry No. 110 at ¶ 188; Docket Entry No. 151-27 at 30). A public observer could reasonably conclude that the statements in Categories 1 and 2 that the pollution issues “ha[d] been remediated” are contradicted by the 2Q10 Form 10-Q’s revelation of continuing pollution issues at well sites “in Susquehanna County.” The fine distinctions on which Cabot relies do not create a genuine mismatch.

The court is persuaded, however, that there is a mismatch between the Guidance Update and the statements in Categories 1 and 2. The Guidance Update announced that Cabot had “adjusted its 2019 production growth guidance to a range of 16 to 18 percent . . . due in large part to a change in the operating plan resulting from a unique opportunity to acquire acreage adjacent to an eight-well pad, allowing the Company to increase the total lateral footage on the pad by approximately 28,000 feet . . . .” (Docket Entry No. 142-14 at 7). The Guidance Update went on to explain that extending the pad would “result in a delay in the wells being placed on production, pushing out the production contribution from this pad to late December or early January.” (*Id.*). The Guidance Update also announced that Cabot “ha[d] updated its 2019 capital budget to a range of \$800 million to \$820 million to reflect the incremental drilling and completion activity on the previously referenced eight-well pad and an increase in drilling activity for the year by four net wells resulting from continued efficiency gains on the Company’s three fully-contracted drilling rigs.” (*Id.*). Finally, the Guidance Update stated that Cabot’s “preliminary 2020 production growth guidance [was] five percent . . . . based on a preliminary capital budget range of \$700 million to \$725 million.” (*Id.* at 7–8). In sum, the Guidance Update told the public that Cabot expected to produce less and spend more in the second half of 2019, and that it expected five percent production growth in 2020. (*See* Docket Entry No. 142-1 at ¶ 32).

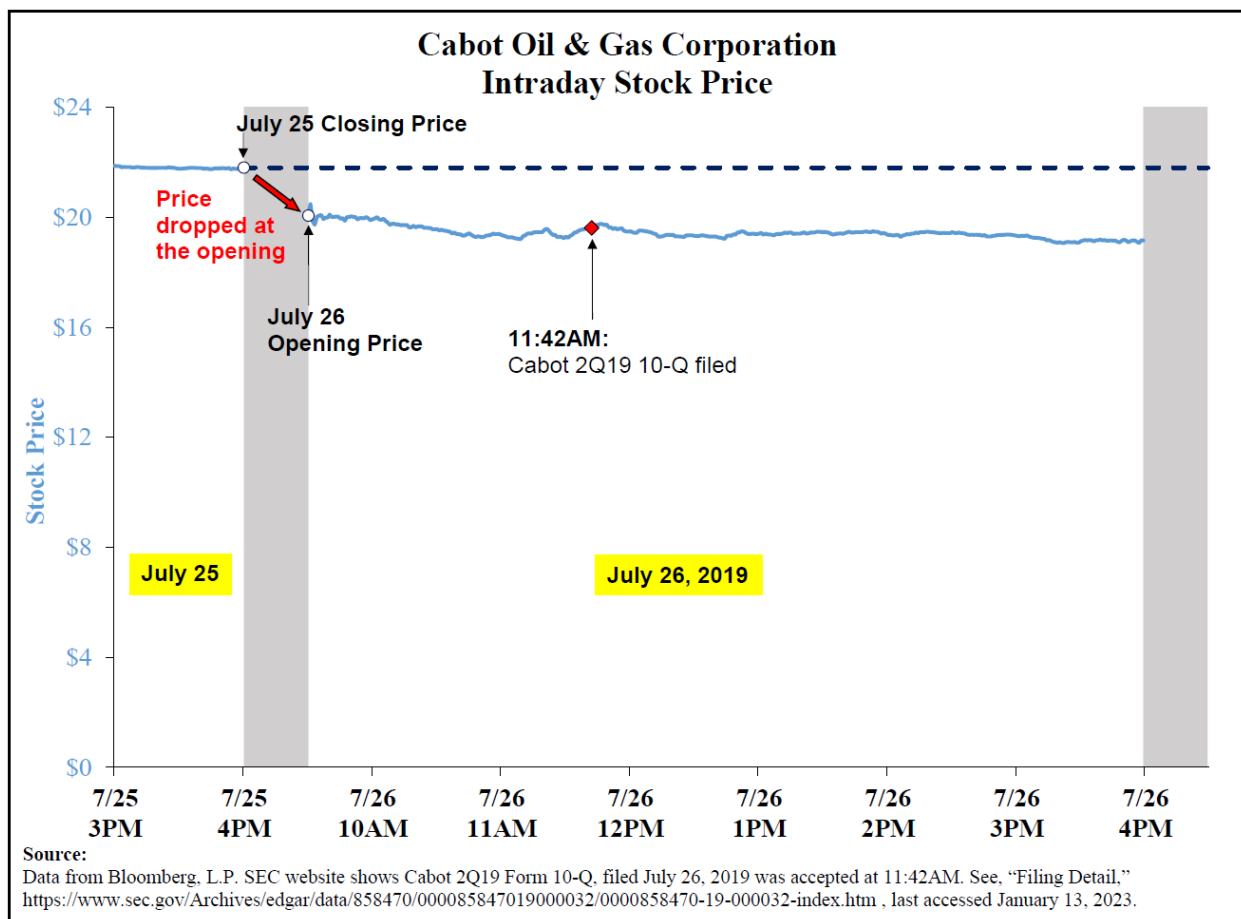
On its face, the Guidance Update had nothing to do with Cabot’s environmental compliance in Susquehanna County. The Plans have produced evidence suggesting that Cabot’s

Case 2:21-cv-00855 Document 1-2 Filed 09/29/23 in TXSD Page 28 of 28  
environmental violations were a hidden cause of its disappointing predicted production, but that would not have been apparent to the public. (See Docket Entry No. 151 at 14–17). It is unlikely that the market would infer from the Production Guidance that Cabot’s previous representations about environmental compliance and remediation in Susquehanna County were false. Accordingly, there is a mismatch between the Production Guidance and the statements in Categories 1–2.

This mismatch, however, is immaterial to the outcome of the class-certification analysis unless Cabot can show that the July 26, 2019, price drop was caused by the Guidance Update and not by the 2Q10 Form 10-Q.<sup>13</sup> Cabot relies on Ms. Allen’s analysis of Cabot’s “intraday” stock price on July 26, 2019. Ms. Allen’s analysis shows that Cabot’s stock price fell around the time the market opened at 9:30 a.m. (Docket Entry No. 142-1 at ¶¶ 31–32). Because Cabot issued the Production Guidance at 6:38 a.m. that day, but did not file its 2Q10 Form 10-Q until 11:42 a.m., Ms. Allen concludes that the price drop was caused by the Production Guidance and not by the Form 10-Q. (*Id.* at ¶ 32).

---

<sup>13</sup> The court disagrees with the Plans that whether the 2Q10 Form 10-Q caused the price drop is a question of “loss causation” inappropriate for inquiry at the class-certification stage. (Docket Entry No. 151 at 11–14). *Goldman* instructed district courts “to take into account *all* record evidence relevant to price impact, regardless whether that evidence overlaps with materiality or any other merits issue.” *Goldman*, 141 S. Ct. at 1961. When causation-related evidence is relevant to the price-impact question, as it is here, courts must consider it.



Cabot’s argument appears persuasive at first blush, but the Plans point out that the stock price fell another 3% between 11:42 a.m. and market close. (Docket Entry No. 151 at 17 (citing Docket Entry No. 151-21 at 15–19)). Cabot has shown that the price drop was caused in part by the Production Guidance; Cabot has not excluded the Form 10-Q as a contributor to the price drop.

Cabot also relies on analyst commentary to show that the Form 10-Q did not contribute to the price drop. According to Cabot’s expert, “[n]one of the 18 analysts following Cabot even mentioned the two [notices of violation] in their reports issued on or after July 26, 2019, indicating that the [notices of violation] were not viewed by analyst[s] as information material to Cabot’s stock price.” (Docket Entry No. 142-1 at ¶ 34). These analysts did, however, treat Cabot’s Guidance Update as material to its stock price. (*Id.* at ¶ 35). Market analysis following a corrective disclosure is relevant to whether the earlier misrepresentation had price impact. *See Goldman*, 77 F.4th at 104. And the market analysis here suggests that the market regarded the Guidance Update



Case 2:21-cv-00855 Document 1-2 Filed 09/29/23 In TXSD Page 22 of 243  
as more significant than the notices of violation disclosed in the 2Q10 Form 10-Q. But the record is inadequate to conclude that the Form 10-Q did not at least contribute to the price drop.

## **2. The Presentment of Charges**

Cabot next argues that there is a mismatch between the Presentment of Charges and the statements in Categories 1 to 3. Cabot argues a mismatch between the Presentment of Charges and the statements in Categories 1 and 2 because Cabot was not charged with violations for its operation of the Stalter Wells, and the statements in Categories 1 and 2 related only to those wells. (Docket Entry No. 142 at 16). Cabot argues that there is a mismatch between the Presentment of Charges and the statements in Category 3 because those statements concern Cabot's conduct at different times. According to Cabot, the charges were for "violations through June 11, 2018," while the 2Q10 Form 10-Q represented only that Cabot had taken "appropriate remediation efforts" as of July 26, 2019. (*Id.* at 16). Put simply, Cabot's argument seems to be that the charge do not contradict the statements in the 2Q10 Form 10-Q because Cabot could have been noncompliant on June 11, 2018, but not on July 26, 2019. Neither argument is persuasive.

First, as noted, the statements in Categories 1 and 2 were not limited to the Stalter Wells. A reasonable observer could infer that the charges against Cabot revealed the falsity of its representations in 2015 and 2016 that it had remediated the wells in Susquehanna County and was "work[ing] with the [Pennsylvania Department] to bring this matter to a close." (Docket Entry No. 110 at ¶ 188).

Cabot's mismatch argument about the Category 3 Statements is based on a misreading of the 2Q10 Form 10-Q. In that Form 10-Q, Cabot represented that it "ha[d] been engaged with the [Pennsylvania Department] in investigating the incidents and ha[d] performed appropriate remediation efforts" since residents in the area raised complaints "in March and June 2017." (Docket Entry No. 151-27 at 30). The charges for violations "through June 11, 2018" is corrective

Case 2:21-cv-00855 Document 12-1 Filed 09/29/23 in TXSD Page 23 of 23  
of Cabot's representation that it had performed "appropriate remediation efforts" between the Spring of 2017 and July 26, 2019. There is no mismatch.

#### IV. Conclusion

The Plans meet the requirements of Rule 23. Cabot has not rebutted the *Basic* presumption by showing no price impact. The Plans' motion for class certification is granted. (Docket Entry No. 134). The court certifies a class consisting of all persons or entities who purchased or otherwise acquired Cabot common stock between February 22, 2016, and June 12, 2020, inclusive, and were damaged thereby. The Plans are appointed as class representatives, and Robbins Geller Rudman & Down LLP and Kessler Topaz Meltzer & Check, LLP are appointed as class counsel.

SIGNED on September 27, 2023, at Houston, Texas.



---

Lee H. Rosenthal  
United States District Judge

# Exhibit 66

**IN THE UNITED STATES DISTRICT COURT  
NORTHERN DISTRICT OF TEXAS  
DALLAS DIVISION**

THE ERICA P. JOHN FUND, INC.,	§	
<i>On Behalf of Itself and All Others Similarly</i>	§	
<i>Situated,</i>	§	
Plaintiffs,	§	
v.	§	No. 3:02-CV-1152-M
	§	
HALLIBURTON COMPANY and	§	
DAVID J. LESAR,	§	
	§	
Defendants.	§	

**MEMORANDUM OPINION AND ORDER**

Before the Court is Plaintiffs’ Motion for Class Certification [Docket Entry #341], Defendants’ Response and Brief on Price Impact [Docket Entry #572], and Plaintiffs’ Price Impact Memorandum [Docket Entry #594]. For the reasons stated herein, the Court **GRANTS in part** Plaintiffs’ Motion for Class Certification, only with respect to the alleged corrective disclosure of December 7, 2001, and **DENIES** Plaintiffs’ Motion for Class Certification as to the other five corrective disclosures on which Plaintiffs rely.

**I. BACKGROUND AND PROCEDURAL HISTORY**

The parties are well-acquainted with the long and winding history of this matter. As a result, for orientation purposes only, the Court will provide the relevant facts from the Supreme Court’s decision in *Halliburton, Co. v. Erica P. John Fund, Inc.*, 134 S. Ct. 2398 (2014) (“*Halliburton II*”):

Erica P. John Fund, Inc. (EPJ Fund), is the lead plaintiff in a putative class action against Halliburton and one of its executives (collectively Halliburton) alleging violations of section 10(b) of the Securities Exchange Act of 1934, 48 Stat. 891, 15 U.S.C. § 78j(b), and Securities and Exchange Commission Rule 10b–5, 17 CFR § 240.10b–5 (2013). According to EPJ Fund . . . Halliburton made a series of misrepresentations regarding its potential liability in asbestos litigation, its expected revenue from certain construction contracts, and the anticipated benefits of its merger with another company—all in an

attempt to inflate the price of its stock. Halliburton subsequently made a number of corrective disclosures, which, EPJ Fund contends, caused the company's stock price to drop and investors to lose money. *Halliburton II*, 134 S. Ct. at 2405–06.

Plaintiffs, represented by the Erica P. John Fund, Inc. (“the Fund”), initially moved to certify a class consisting of all investors who bought Halliburton common stock between June 3, 1999 and December 7, 2001.<sup>1</sup> This Court found that the proposed class met all of the prerequisites of Federal Rule of Civil Procedure 23(a)—numerosity, common questions of law and fact, typicality, superiority, and adequacy. However, the Court denied class certification because Fifth Circuit precedent required Plaintiffs to prove “loss causation” to invoke the fraud-on-the-market presumption of *Basic v. Levinson*, 485 U.S. 224, 243 (1988), and the Court concluded the Fund had not met its burden of proof to do so.<sup>2</sup> The Fifth Circuit affirmed on that ground.<sup>3</sup> The Supreme Court subsequently vacated the judgment, holding that loss causation “addresses a matter different from whether an investor relied on a misrepresentation, presumptively or otherwise, when buying or selling a stock,” and that loss causation need not be shown at the class certification stage. *Erica P. John Fund, Inc. v. Halliburton Co.*, 131 S. Ct. 2179, 2185–86 (2011) (“*Halliburton I*”). The case was remanded to this Court to address any “further arguments against class certification” preserved by Halliburton. *Id.* at 2187.

Halliburton argued on remand that the evidence it had presented to disprove loss causation also demonstrated that none of the alleged misrepresentations actually impacted Halliburton's stock price, *i.e.*, there was a lack of “price impact,” and, therefore, Halliburton had

---

<sup>1</sup> *Archdiocese of Milwaukee Supporting Fund, Inc. v. Halliburton Co.*, No. 3:02-CV-1152-M, 2008 WL 4791492, at \*1 (N.D. Tex. Nov. 4, 2008) *aff'd*, 597 F.3d 330 (5th Cir. 2010) *vacated and remanded sub nom.*, *Erica P. John Fund, Inc. v. Halliburton Co.*, 131 S. Ct. 2179 (2011).

<sup>2</sup> *Id.*

<sup>3</sup> *Archdiocese of Milwaukee Supporting Fund, Inc. v. Halliburton Co.*, 597 F.3d 330 (5th Cir. 2010) *vacated and remanded sub nom.*, *Erica P. John Fund, Inc. v. Halliburton Co.*, 131 S. Ct. 2179 (2011).

rebutted the *Basic* presumption that the Fund and other members of the class relied on the misrepresentations when they bought and sold Halliburton's stock at the market price.<sup>4</sup>

Halliburton argued the Fund and other putative class members would have to prove reliance on an individual basis, thereby causing individual issues to predominate over common issues.<sup>5</sup> This Court rejected that argument, and the Fifth Circuit again affirmed, holding that evidence of the absence of price impact to rebut the *Basic* presumption is not relevant to predominance under Rule 23(b)(3), but can be admitted at trial.<sup>6</sup>

In *Halliburton II*, the Supreme Court reversed, holding that Halliburton could introduce evidence of a lack of price impact at the class certification stage to show the absence of predominance. *Halliburton II*, 134 S. Ct. at 2414–17. The Supreme Court again vacated the judgment of the Fifth Circuit and remanded the case to this Court for further proceedings. *Id.* at 2417.

After *Halliburton II* was issued, this Court ordered the Fund and Halliburton to provide additional briefing on price impact as it relates to class certification.<sup>7</sup> Each party submitted an expert report and additional briefing, and the Court held an evidentiary hearing.<sup>8</sup> Both the Fund and Halliburton filed *Daubert* motions to exclude each other's experts, which the Court denied,

---

<sup>4</sup> *Archdiocese of Milwaukee Supporting Fund, Inc. v. Halliburton Co.*, No. 3:02-CV-1152-M, 2012 WL 565997 (N.D. Tex. Jan. 27, 2012) *aff'd sub nom.*, *Erica P. John Fund, Inc. v. Halliburton Co.*, 718 F.3d 423 (5th Cir. 2013) *vacated and remanded*, 134 S. Ct. 2398 (2014) ("*Halliburton II*").

<sup>5</sup> *Id.*

<sup>6</sup> *Erica P. John Fund, Inc. v. Halliburton Co.*, 718 F.3d 423 (5th Cir. 2013) *cert. granted*, 134 S. Ct. 636 (2013) *vacated and remanded*, 134 S. Ct. 2398 (2014).

<sup>7</sup> Dkt. No. 568.

<sup>8</sup> Dkt. No. 572; Dkt. No. 590; Dkt. No. 598.

having determined that the parties' arguments about the reliability of the experts' methods were inextricably intertwined with the parties' merits arguments on price impact.<sup>9</sup>

Plaintiffs now seek to certify a class commencing on July 22, 1999, a later date than that originally requested. It was on July 22, 1999 that Halliburton announced its Second Quarter 1999 results and held an earnings call. Plaintiffs claim the class period should end on December 7, 2001, when Halliburton announced the verdict in Maryland against Halliburton's subsidiary, Dresser.<sup>10</sup> Plaintiffs seek to certify a class for only asbestos and accounting claims, not for claims relating to the Dresser merger.<sup>11</sup>

## II. ANALYSIS

### A. Issues Before the Court

Section 10(b) of the Exchange Act of 1934, and Rule 10b-5, which prohibit making material misstatements or omissions in connection with the purchase or sale of a security, are enforced by an implied private cause of action. In order to prevail and recover damages, a plaintiff must prove ““(1) a material misrepresentation or omission by the defendant; (2) scienter; (3) a connection between the misrepresentation or omission and the purchase or sale of a security; (4) reliance upon the misrepresentation or omission; (5) economic loss; and (6) loss causation.”” *Amgen, Inc. v. Connecticut Retirement Plans and Trust Funds*, 133 S. Ct. 1184, 1192 (2013).

The element of reliance “ensures that that there is a proper connection between a defendant's misrepresentation and a plaintiff's injury.” *Halliburton I*, 131 S. Ct. at 2184–85 (citing *Basic*, 485 U.S. at 243). “The traditional (and most direct) way a plaintiff can

---

<sup>9</sup> Dkt. No. 598; Hr'g Tr. at 247:6–12.

<sup>10</sup> Dkt. No. 590 at 2 n. 1.

<sup>11</sup> *Id.*



demonstrate reliance is by showing that he was aware of a company's statement and engaged in a relevant transaction—*e.g.*, purchasing common stock—based on that specific misrepresentation.” *Id.* at 2185. However, the Supreme Court in *Basic* explained that requiring “such direct proof of reliance ‘would place an unnecessarily unrealistic evidentiary burden on the Rule 10b-5 plaintiff who has traded on an impersonal market.’” *Halliburton II*, 134 S. Ct. at 2407 (quoting *Basic*, 485 U.S. at 245). Furthermore, the Court in *Basic* recognized that “[r]equiring proof of individualized reliance’ from every securities fraud plaintiff ‘effectively would . . . prevent [ ] [plaintiffs] from proceeding with a class action’ in Rule 10b-5 suits” because “individual issues then would . . . overwhelm [ ] the common ones,” thus precluding certification under Rule 23(b)(3). *Id.* (quoting *Basic*, 485 U.S. at 242).

Given the difficulties the reliance element posed for securities fraud plaintiffs, the Court in *Basic* held that such plaintiffs may, in limited circumstances, satisfy the reliance element of a Rule 10b-5 suit by invoking a rebuttable presumption of reliance. *Halliburton II*, 134 S. Ct. at 2408. Based on the “fraud-on-the-market” theory, a plaintiff must meet the following elements to invoke the *Basic* presumption: “(1) that the alleged misrepresentations were publicly known, (2) that they were material, (3) that the stock traded in an efficient market, and (4) that the plaintiff traded the stock between the time the misrepresentations were made and when the truth was revealed.” *Id.* (citations omitted). The Court in *Basic* further explained that this presumption may be rebutted by “[a]ny showing that severs the link between the alleged misrepresentation and either the price received (or paid) by the plaintiff, or his decision to trade at a fair market price.” *Basic*, 485 U.S. at 248. Therefore, “if a defendant could show that the alleged misrepresentation did not, for whatever reason, actually affect the market price, or that a plaintiff would have bought or sold the stock even had he been aware that the stock’s price was

tainted by fraud,” the presumption of reliance would have been rebutted. *Halliburton II*, 134 S. Ct. at 2408 (citing *Basic*, 485 U.S. at 248–49).

The Court has already found, and still finds, that the Fund has met the certification requirements of numerosity, commonality, typicality, superiority, and adequacy under Federal Rule of Civil Procedure 23. *See Halliburton II*, 134 S. Ct. at 2406. Those findings remain undisturbed. *See Archdiocese of Milwaukee Supporting Fund, Inc. v. Halliburton Co.*, No. 3:02-1152-M, 2012 WL 565997 (N.D. Tex. Jan. 27, 2012), *vacated on other grounds sub nom.*, *Halliburton II*, 134 S. Ct. 2398 (2014). The only issue before the Court on class certification is predominance under Federal Rule of Civil Procedure 23(b), which requires that “questions of law or fact common to class members predominate over any questions affecting only individual class members.” Fed. R. Civ. P. 23(b).

To overcome the difficulties of showing reliance on a class-wide basis, the Fund must demonstrate that the *Basic* presumption applies, by showing that Halliburton’s misrepresentations were publicly known and material, that Halliburton’s stock traded in an efficient market, and that Plaintiffs traded the stock between the times the alleged misrepresentations were made and when the relevant truths were revealed. Prior to *Halliburton II*, if it did so, this Court would have treated its inquiry as then at an end. However, in *Halliburton II*, the Supreme Court clarified that securities fraud defendants may rebut the *Basic* presumption at the class certification stage by presenting evidence of lack of price impact. *Halliburton II*, 134 S. Ct. at 2417. Thus, Halliburton now has the opportunity to rebut the presumption “with evidence that the asserted misrepresentation (or its correction) did not affect the market price.” *Id.* at 2414.

Accordingly, the parties have submitted event studies, *i.e.*, regression analyses, to show that Halliburton's stock price was, or was not, affected on days when an alleged misrepresentation or corrective disclosure reached the market. *See id.* at 2415 (explaining that event studies may show both market efficiency and a lack of price impact). The parties also raise two threshold legal issues the Court will address before analyzing the parties' event studies—(1) who has the burden of production and persuasion; and (2) whether the Court should, as part of the price impact inquiry, rule as a matter of law that particular disclosures are corrective.

### **B. Burdens of Production and Persuasion**

The Court is of the opinion, and the parties seem to agree, that the placement of the burdens of production and persuasion in this case does not alter the Court's decision on the merits.<sup>12</sup> The Supreme Court did not state expressly in *Halliburton II* whether plaintiffs or defendants must carry the burden of persuasion to show price impact or lack thereof, but based on the Court's analysis of the Supreme Court's decision in *Halliburton II*, and decisions by other district courts since *Halliburton II*, the Court finds the burdens of production and persuasion to show lack of price impact are properly placed on Halliburton.

Halliburton argues that neither the majority opinion nor the concurrence in *Halliburton II* expressly stated that a defendant must prove a lack of price impact. Halliburton also contends that placing the burden of persuasion on the Fund is consistent with Federal Rule of Evidence 301, which states:

In a civil case, unless a federal statute or these rules provide otherwise, the party against whom a presumption is directed has the burden of producing evidence to rebut the presumption. But this rule does not shift the burden of persuasion, which remains on the party who had it originally.

---

<sup>12</sup> *See* Hr'g Tr. at 7:18–22, 17:6–12.

Fed. R. Evid. 301. Thus, Halliburton argues that under Rule 301, which was cited by the Supreme Court in *Basic*, the defendant only has the burden of “producing evidence to rebut the presumption.” 485 U.S. at 245 (stating that presumptions are useful devices for allocating the burdens of proof between parties). Halliburton contends that the Fund has the burden of showing reliance, and the *Basic* presumption allows it to satisfy that burden if the Fund can establish the facts that give rise to the presumption. However, once Halliburton produces evidence to rebut the presumption, the presumption is rebutted and the case can no longer move forward on a class basis. Dkt. No. 572 at 3; *see also* Merrit B. Fox, *Halliburton II: It All Depends on What Defendants Need to Show to Establish No Price Impact*, 70 Bus. Law 437, 457 n. 47 (2014-15) (summarizing the Fed. R. Evid. 301 argument). Halliburton further argues that plaintiffs have the burden of proving predominance under Rule 23(b)(3), and because price impact has everything to do with the issue of predominance, the burdens associated with the issue should fall on the Fund. According to Halliburton, once it rebuts the presumption, a class may not be certified unless the Fund proves price impact.

The Fund responds that *Halliburton II* places both the burden of production and persuasion on Halliburton, because the Supreme Court reaffirmed *Basic*’s presumption of reliance, but held defendants could rebut the presumption at the class certification stage by carrying the burdens of production and persuasion.<sup>13</sup> In their concurrence, Justices Sotomayor, Ginsburg, and Breyer stated that the “Court recognizes that it is incumbent upon the defendant to show the absence of price impact.” 134 S. Ct. at 2417 (citing majority opinion at 2413–2414). Plaintiffs argue that prior Fifth Circuit precedent remains unchanged and, therefore, the burden

---

<sup>13</sup> *See* Hr’g Tr. at 16:25–17:1.

of persuasion is on Halliburton. *See Abell v. Potomac Ins. Co.*, 858 F.2d 1104, 1118 (5th Cir. 1988), *vacated in part on other grounds sub nom. Fryar v. Abell*, 492 U.S. 914 (1989).

The Fund further argues that, despite Halliburton’s invocation of Rule 301 in its briefing to the Supreme Court, the Court did not reference Rule 301 in *Halliburton II*. Moreover, the Fund claims that the fraud-on-the-market presumption is “a substantive doctrine of federal securities law” and is not a creation of Rule 301. *See Halliburton II*, 134 S. Ct. at 2411 (citing *Amgen*, 133 S. Ct. at 1193). Finally, the Fund contends that relieving Halliburton of the burden of persuasion would eviscerate the *Basic* presumption because Halliburton could arguably satisfy its burden merely by having an expert opine that price impact was absent. Shifting the burden would require the Fund to prove price impact directly at class certification—a proposal the Supreme Court said would radically alter the reliance showing.

In *Halliburton II*, the Court saw no reason to “artificially limit the [price impact] inquiry at the certification stage to indirect evidence,” and authorized defendants to seek to defeat the *Basic* presumption at the class certification stage through direct as well as indirect price impact evidence. 134 S. Ct. at 2417. By requiring plaintiffs to carry the burden of persuasion to show price impact at the class certification stage, this Court would, in effect, be requiring the Fund to prove price impact directly, a proposition the Supreme Court refused to adopt. Indeed, as the concurrence noted, requiring direct proof at the class certification stage would have a radical impact on 10b-5 class actions. *See Halliburton II*, 134 S. Ct. at 2417 (Ginsburg, J., Sotomayor, J., Breyer, J., concurring) (“The Court’s judgment . . . should impose no heavy toll on securities-fraud plaintiffs with tenable claims” because “it is incumbent upon the defendant to show the absence of price impact.”).

In *Aranaz v. Catalyst Pharmaceutical Partners, Inc.*, the court stated that defendants had the burden of persuasion to show an absence of price impact at the class certification stage, and concluded they failed to carry that burden, where there was a clear and drastic spike in price and an equally drastic decline following the misrepresentation and correction, respectively. 302 F.R.D. 657, 673 (S.D. Fla. 2014). In *McIntire v. China MediaExpress Holdings, Inc.*, consistent with existing precedent in the Second Circuit, the court stated that the defendant bore the burden at the certification stage to prove a lack of price impact. 38 F. Supp. 3d 415, 434 (S.D.N.Y. 2014) (citing *Halliburton II*, 134 S. Ct. at 2417 (Ginsburg, J., concurring)). In *Wallace v. IntraLinks*, the court also found that the defendants bore the burden to show a lack of price impact. 302 F.R.D. 310, 317 (S.D.N.Y. 2014) (citing *McIntire*, 38 F. Supp. 3d at 434).

Halliburton relies on Rule 301 to support its argument that it should bear only the burden of production. First, it is worth noting that, in *Basic*, the Court cited Rule 301 merely to illustrate the usefulness of presumptions in allocating the burden of proof. The *Basic* presumption's consistency with Rule 301 was one of many reasons the Court cited to support the presumption's creation. *See Basic*, 485 U.S. at 245 (justifying the presumption as being supported by common sense, probability, and consistency with congressional policy in enacting the 1934 Securities Exchange Act). Moreover, the very nature of the fraud-on-the-market presumption makes it difficult to apply Rule 301 in a direct manner:

Rule 301 requires the party against whom a presumption is directed (the second party) to produce evidence suggesting the non-existence of the basic facts needed to establish the presumption. [Rule 301] seems to contemplate, though, that this evidence need only be sufficient enough to meet the burden of going forward. At this point the presumption disappears and the party that sought to invoke the presumption (the first party), without the aid of the presumption, has the burden of persuasion as to the fact that the presumption presumed. Behind this seemingly harsh rule appears to be a hidden assumption: the facts that need to be established to give rise to the presumption are probative as to the existence of the facts that the presumption presumes. So, while the

first party no longer has the benefit of the presumption, it still has the benefit of the probative value of the evidence that it produced to originally give rise to the presumption.

Fox, *supra* p. 8, at 457–58. However, as Professor Fox explains, the fraud-on-the-market presumption is atypical, and as a result, does not neatly fit into the Rule 301 framework:

The basic facts that the plaintiff needs to establish to give rise to [the fraud-on-the-market presumption]—the materiality of the misstatement and the efficiency of the market for the issuer’s shares—are not probative to whether plaintiffs actually relied on the misstatement in the traditional sense. In other words, these basic facts do not help demonstrate that but for the misstatement, the plaintiffs would not have bought their shares. Rather, they are probative to whether the misstatement affected price.

*Id.* at 458. The Court in *Basic* essentially packaged a new cause of action as a presumption:

[I]f the plaintiffs establish the specified facts giving rise to [the fraud-on-the-market presumption]—materiality and market efficiency—the plaintiff need not prove something that has been traditionally required in fraud-based damage actions, i.e. , that but for the misstatement, each plaintiff would not have purchased her shares. Unlike the usual presumption, however, the facts needed to establish the fraud-on-the-market presumption are entirely unrelated to the likelihood that the fact presumed by the presumption actually existed.

*Id.* As Professor Fox explains, a literal application of Rule 301 to the fraud-on-the-market presumption in a class certification hearing would allow defendants to preclude class certification by merely putting on a reputable expert that can opine with 95% confidence that a corrective disclosure had no effect on price. *Id.* at 458–59. According to Halliburton’s position on Rule 301, the Fund would then be forced to move forward and prove reliance without the aid of the presumption, which would doom the class on predominance grounds. The Fund would not be afforded an opportunity to salvage the class by producing its own reputable expert to challenge Halliburton’s. The Court finds that the Supreme Court would not have modified the fraud-on-the-market presumption so substantially without explicitly saying so.

Thus, to the extent it matters with respect to any of the conclusions reached below, the Court finds that both the burden of production and the burden of persuasion are properly placed



on Halliburton. In other words, Halliburton must ultimately persuade the Court that its expert's event studies are more probative of price impact than the Fund's expert's event studies.

### C. Corrective Disclosures

Halliburton raises a threshold legal issue in its briefing on lack of price impact. It argues that each of the alleged corrective disclosures were not, in fact corrective, and therefore, Halliburton has rebutted the fraud-on-the-market presumption by showing a severance of the “link between the alleged misrepresentation and . . . the price received (or paid) by the plaintiff” because “the basis for the finding that the fraud had been transmitted through the market price [is] gone.” *See Halliburton II*, 134 S. Ct. at 2415–16. Halliburton claims that there is evidence of no price impact when the alleged misrepresentation disclosed information already known by the market; said another way, a defendant rebuts the presumption by showing there was no *correction* that affected the market price.<sup>14</sup>

Accordingly, Halliburton argues that this Court's prior findings and the Fifth Circuit's findings relating to loss causation remain intact, and both Courts found that the alleged corrective disclosures were not, as a matter of law, corrective of the alleged misrepresentations.

The Fund counters that the class certification stage is not an appropriate procedural stage for the Court to rule on whether the disclosures at issue were corrective, and that this legal issue is more properly addressed at the pleading stage or the later summary judgment stage of proceedings.<sup>15</sup> Rather, the Fund maintains that this Court's inquiry should be limited to price impact—assuming the disclosure was corrective of an earlier fraud, did Halliburton's stock price move on the day of the disclosure?<sup>16</sup>

---

<sup>14</sup> Dkt. No. 572 at 2.

<sup>15</sup> Hr'g Tr. at 17:18–18:3.

<sup>16</sup> *Id.* at 18:7–8.

Based on the Supreme Court’s discussion in *Halliburton I*, *Amgen*, and *Halliburton II*, the Court finds that class certification is not the proper procedural stage for the Court to determine, as a matter of law, whether the relevant disclosures were corrective. Furthermore, the Court finds that Halliburton’s arguments regarding whether the disclosures were corrective are, in effect, a veiled attempt to assert the “truth on the market” defense, which pertains to materiality and is not properly before the Court at this stage of the proceedings. *See Aranaz*, 302 F.R.D. at 671 (“[F]or purposes of determining at this early stage of litigation whether the alleged misrepresentation had any impact on the price of [the defendant’s] stock, the Court must disregard evidence that the truth was known to the public.”) (citing *Amgen*, 133 S. Ct. at 1203).<sup>17</sup>

In *Halliburton I*, the Supreme Court held that loss causation is not a precondition for invoking *Basic*’s rebuttable presumption of reliance. 131 S. Ct. at 2185–86. The Court explained that loss causation “addresses a matter different from whether an investor relied on a misrepresentation, presumptively or otherwise, when buying or selling a stock.” *Id.* at 2186. The Court described the reliance element in a 10b-5 action as “transaction causation,” as distinguished from loss causation, which means the Court “focus[es] on the facts surrounding the investor’s decision to engage in the transaction.” *Id.* (citing *Dura Pharmaceuticals, Inc. v. Broudo*, 544 U.S. 336, 342 (2005)). In holding that loss causation need not be proved at the certification stage, the Court reasoned that “[l]oss causation has no logical connection to the facts necessary to establish the efficient market predicate to the fraud-on-the-market theory.” *Id.* The

---

<sup>17</sup> For example, Halliburton argues that the December 21, 2000 disclosure that it was taking a \$120 million charge, in part due to losses on construction projects, was not corrective because Halliburton had previously disclosed the risk of cost overruns, and the market knew that such construction projects carried risk. Allen Rep. at 55 ¶ 116. In other words, Halliburton argues the market was *already aware* of what was disclosed on December 21, 2000. However, this is an argument more properly associated with materiality. *See Aranaz*, 302 F.R.D. at 671.

Court found that requiring securities fraud plaintiffs to prove loss causation at the class certification stage would contradict *Basic*'s fundamental premise that an investor presumptively relies on a misrepresentation when it was incorporated in the market price at the time of the transaction. *Id.*

In *Amgen*, the Supreme Court held that securities fraud plaintiffs need not prove materiality at the class certification stage, reasoning that "the question of materiality is common to the class, and [ ] a failure of proof on that issue would not result in questions 'affecting only individual members' predominating . . . ." *Amgen*, 133 S. Ct. at 1197. The Court found that plaintiffs' inability to prove materiality would not create a "fatal dissimilarity" among class members and make the class-action mechanism inefficient or unfair. *Id.* Indeed, the Court found that materiality could be proven through evidence common to the class, and insufficient or absent evidence on materiality would "end the case for one and for all" because "no claim would remain in which individual reliance issues could potentially predominate." *Id.* at 1195. The Court found that *Amgen*, in seeking to challenge materiality, was seeking to disprove an element of a Rule 10b-5 cause of action, which is more properly dealt with at trial or on a motion for summary judgment. *Id.*

Finally, in *Halliburton II*, in rejecting the Fund's argument that *Amgen*'s reasoning should apply to preclude *Halliburton* from introducing evidence of lack of price impact at the certification stage, the Court found that "price impact differs from materiality in a crucial respect." 134 S. Ct. at 2416. According to the Court, "materiality is a discrete issue that can be resolved in isolation from the other prerequisites" and "can be wholly confined to the merits stage." *Id.* In contrast, price impact goes to *Basic*'s fundamental premise—"the fact that a misrepresentation 'was reflected in the market price at the time of [the] transaction.'" *Id.* (citing

*Halliburton I*, 131 S. Ct. at 2186). Because “publicity” and “market efficiency” are merely prerequisites for an indirect showing of price impact, evidence of which will already be before a court determining whether the *Basic* presumption applies, the Supreme Court held that defendants should be able to present evidence to rebut the *Basic* presumption at the certification stage. *Id.* at 2416–17.

This Court holds that *Amgen* and *Halliburton I* strongly suggest that the issue of whether disclosures are corrective is not a proper inquiry at the certification stage. *Basic* presupposes that a *misrepresentation* is reflected in the market price at the time of the transaction. *See Halliburton II*, 134 S. Ct. at 2416. Thus, at this stage of the proceedings, the Court concludes that the asserted misrepresentations were, in fact, misrepresentations, and assumes that the asserted corrective disclosures were corrective of the alleged misrepresentations. To hold otherwise would require the Court to pass judgment on the merits of the allegations after the dismissal stage and before summary judgment—in effect, giving a third bite at the apple to *Halliburton*. While it may be true that a finding that a particular disclosure was not corrective as a matter of law would “sever the link between the alleged misrepresentation and . . . the price received (or paid) by the plaintiff . . .,” the Court is unable to unravel such a finding from the materiality inquiry. *See Halliburton II*, 134 S. Ct. at 2415–16. Finally, a finding at this stage that a disclosure was not corrective will not result in a predominance problem, *i.e.*, a “fatal dissimilarity” that causes individual questions of law and fact to predominate over common questions. *See Amgen*, 133 S. Ct. at 1195–96. In other words, if *Halliburton* were to successfully persuade the Court at summary judgment that a particular disclosure was not corrective, it would end this controversy altogether. *See Aranz*, 302 F.R.D. at 671.

#### D. Price Impact

Fraud on the market securities litigation typically focuses on a price change at the time of a corrective disclosure. Fox, *supra* p. 8, at 441. If a particular disclosure causes the stock price to decline at the time of disclosure, then the misrepresentation must have made the price higher than it would have otherwise been without the misrepresentation. *Id.* Measuring price change at the time of the corrective disclosure, rather than at the time of the corresponding misrepresentation, allows for the fact that many alleged misrepresentations conceal a truth. *Id.* Thus, the misrepresentation will not have changed the share price at the time it was made. *Id.*

To show that a corrective disclosure had a negative impact on a company's share price, courts generally require a party's expert to testify based on an event study that meets the 95% confidence standard, which means "one can reject with 95% confidence the null hypothesis that the corrective disclosure had no impact on price." *Id.* at 442 n. 17. An event study is generally comprised of two parts: (1) a calculation of the market-adjusted price change in the issuer's share price at the time the corrective disclosure became public, *i.e.*, the "difference between the observed price change [in Halliburton's stock price] and what the simultaneous change in overall stock market prices predicts would have been [Halliburton's] price change"; and (2) a determination of whether the corrective disclosure is among the [Halliburton-related] news that affected the price on the date the disclosure became public, *i.e.*, "ask[] how unusual it would be that the observed-market-adjusted price change is due solely to the day's other bits of [Halliburton-related] news and thus not in any part due to the corrective disclosure . . . by comparing the magnitude of the market-adjusted change in [Halliburton's] share price on the date of the corrective disclosure with the historical record of the daily, market-adjusted ups and downs in [Halliburton's] share price." *See id.* at 443.

## 1. Expert Findings

The determination of whether lack of price impact has been shown largely turns on the competing methodologies of the parties' experts.

Halliburton's expert, Lucy Allen ("Allen"), previously submitted reports on market efficiency and loss causation, as did the Fund's prior expert, Jane Nettesheim ("Nettesheim"). The Fund obtained a new expert, Chad Coffman ("Coffman"), to conduct its event studies for price impact testimony.

Allen reviewed the Fund's Complaint, its prior expert reports, and other pleadings to ascertain when the alleged misrepresentations were made, what was alleged to be false, when the truth was allegedly revealed to the market, and what was alleged to be corrective in the asserted corrective disclosures. Allen Rep. at 9 ¶ 15. Allen identified twenty-five dates on which misrepresentations occurred and thirteen dates on which corrective disclosures occurred. *Id.* ¶ 16. Because three of the dates had both a corrective disclosure and an alleged misrepresentation, Allen analyzed thirty-five separate dates. *Id.*

Allen then collected publicly available information about Halliburton, its peers, and issues that affected those firms' stock prices during the class period, including press releases, conference calls, SEC filings, analyst reports, news stories, reports by credit rating agencies, trade publications, and data on expected future volatility of stock prices. *Id.* at 9–11 ¶ 17. Allen reviewed the collected information and commentary to determine what market factors affected Halliburton's stock price, what information was publicly known at what time, how analysts interpreted publicly known information about the alleged misrepresentations, and what analysts considered important and new in the alleged corrective disclosures. *Id.* at 11 ¶ 18.

Allen developed a market model and performed an event study to determine whether there was statistically significant price movement on the dates of the alleged misrepresentations and corrective disclosures. *Id.* ¶ 19. Event studies are used by academics to determine whether and how stock prices respond to new information, and they typically measure movement in a stock price after an event, adjusting for movement in the overall market and/or industry. *Id.* As discussed below, Allen selected specific indices to adjust for movements in the energy services and energy and construction industries. *Id.* at 13 ¶ 21. Her regression analysis estimated the relationship between Halliburton’s daily stock returns and the daily returns of the two industry indices she chose. The results of Allen’s regression analysis and the return of the industry indices were utilized to predict Halliburton’s price movement. The difference between Halliburton’s predicted return and its actual stock return constituted the portion of Halliburton stock price movement not explained by contemporaneous movements in Halliburton’s industries. *Id.* Allen then tested the statistical significance of Halliburton’s excess stock price movement, *i.e.*, price reaction, after each of the misrepresentations, corrective disclosures, and other relevant dates. *Id.* When analyzing the statistical significance of a price reaction to an event, Allen used the commonly-applied 95% confidence level, which means there is a 5% chance of finding a statistically significant price reaction even when there is company-specific news on the market and the price moves according to normal daily fluctuations.<sup>18</sup> *Id.* ¶ 22. As is discussed below, Allen also applied a Bonferroni multiple comparison adjustment, the propriety of which is disputed by the Fund. *Id.* at 13–14 ¶¶ 23–25. Finally, Allen used the information and analyses described to examine whether the alleged misrepresentations impacted Halliburton’s stock price.

---

<sup>18</sup> In other words, “for every 100 price reactions analyzed at the 95% confidence level, on average, 5 will be statistically significant for no reason other than the normal daily variation in the stock price.” Allen Rep. at 13 ¶ 22.



*Id.* at 15 ¶ 27. Ultimately, Allen found no price impact after any of the alleged misrepresentations or corrective disclosures, with the exception of December 7, 2001, about which Allen states there was no price reaction *as to the alleged misrepresentation*, which the Court interprets to mean that the price reaction was caused by factors other than Halliburton's disclosure of an adverse asbestos verdict.

In his rebuttal report, the Fund's expert, Chad Coffman, argued that there are six relevant events in this case on which to evaluate price impact: an accounting corrective disclosure on December 21, 2000, and asbestos corrective disclosures on June 28, 2001, August 9, 2001, October 30, 2001, December 4, 2001, and December 7, 2001. Coffman Rep. at 2–3 ¶ 8. Coffman claimed that his event study and analysis show that the market responded significantly to each of these six events. *Id.* at 3 ¶ 9. Coffman presumed that Plaintiffs' allegations are true—that Halliburton made material misrepresentations or omissions, with scienter, regarding its asbestos liability and its accounting on fixed-price contracts. *Id.* at 12 ¶ 16. Coffman objects to several of the methods employed by Allen in her report, particularly her use of a multiple comparison adjustment and the choice of industry indices she used as bases to compare Halliburton's stock price movement.

Allen found no evidence of price impact as to the thirteen corrective disclosures identified by the Fund's prior expert, Jane Nettesheim. Hr'g Tr. at 8:7-11. Furthermore, according to Allen, Nettesheim's method of selecting the thirteen corrective disclosures was flawed, because Nettesheim first looked at every trading day during the 633 day class period, performed an event study testing Halliburton's stock price on each day, and then looked for Halliburton-specific news on the days on which Halliburton's stock had a significant price reaction. Allen Rep. at 8 ¶ 13. If the news disclosed on the date of a significant price reaction

related to the allegations in the Complaint, Nettesheim treated it as a corrective disclosure. Hr’g Tr. at 8:12-24. Halliburton argues that the six dates now identified by the Fund are merely a subset of the thirteen dates identified by Nettesheim, dates Halliburton alleges were derived from Nettesheim’s original flawed methodology.<sup>19</sup> *Id.* at 32:10-12, 21-25, 33:6-9. The Fund responds that Nettesheim denied coming up with her dates in the way Allen argues, and that the Fund’s new expert, Mr. Coffman, maintains the dates were selected because on those dates there was news related to the allegations in the Complaint, not because there was statistically significant price movement on those dates. *Id.* at 20:5-7; Coffman Rep. at 27 ¶ 48.

There is insufficient evidence before the Court to conclude that Nettesheim’s methodology was flawed in the way Halliburton alleges. Nettesheim prepared reports regarding market efficiency and loss causation, not price impact. Here, Halliburton must show lack of price impact by showing that Allen’s findings are more persuasive than Coffman’s.

According to Allen, there are three primary differences between her methodology and Coffman’s: (1) Coffman used an additional industry index to adjust for movements in Halliburton’s specific industries; (2) Coffman made a multiple comparison adjustment only for six dates, rather than for the thirty-five separate dates alleged as either misrepresentations or corrective disclosures in the Complaint; and (3) Coffman applied a Holm-Bonferroni multiple-comparison adjustment.

For his part, Coffman argues there are two fundamental flaws to Allen’s approach: (1) she did not adequately control for stock price movements in Halliburton’s specific industries, and Coffman claims he dramatically increased the explanatory power of Allen’s event study by

---

<sup>19</sup> Allen explained, “[Coffman’s] testing the height of people and not taking into account the fact that Nettesheim already started and handed him a list of basketball players.” Hr’g Tr. at 43:3-5.

creating an additional peer index; and (2) Allen's multiple comparison adjustment is novel, unnecessary, and yields erroneous results, because it in effect requires greater than a 99% confidence level for a stock price movement on a particular day to be considered statistically significant. If one were to use a multiple comparison adjustment, Coffman argues a Holm-Bonferroni adjustment would be more appropriate than a Bonferroni adjustment because it lowers the risk of false negatives—in other words, it lessens the likelihood of finding no price reaction when there actually is a price reaction. Coffman Rep. at 4 ¶ 9(ii)-(iii).

## 2. Control Group

According to Allen, the fundamental flaw in Coffman's methodology was his construction and use of the control period. The control period is used to estimate the relationship between Halliburton and industry indices. An event study should take into account how Halliburton's stock moves relative to industry indices during the control period, and then that relationship should be used to predict on a test date whether Halliburton's stock price movement is significantly different from what would have been predicted under normal circumstances. Hr'g Tr. at 46:8-17.

Allen considered Halliburton's relationship to the industry indices by looking at the dates during the class period, excluding the thirty-five test dates Nettesheim claimed had a misrepresentation or corrective disclosure. Hr'g Tr. at 44:13-15. Allen concluded the other dates during the class period constituted the control group, to which she could compare the thirty-five test dates. *Id.* at 44:15-19. In contrast, Coffman did not test the thirty-five dates, instead testing only six dates; however, the thirty-five dates were also not in his control group, which, according to Allen, made his methodology internally inconsistent. *Id.* at 44:20-22.

Coffman also used the longer class period originally claimed, whereas Allen used the newer, shorter class period. *Id.* at 47:4-14.

To cure what she claims were Coffman's internal inconsistencies, Allen made two adjustments to Coffman's model: (1) she included the thirty-five test dates in his control period; and (2) she applied his methodology to the new class period. Allen also adopted all three changes Coffman suggested to her model, after which she found only two dates to be statistically significant—August 9, 2001 and December 7, 2001. *Id.* at 48:7-10, 17-21.

The Court is persuaded that Allen's adjustments to Coffman's model were appropriate to achieve internal consistency. The Fund provided no evidence, nor argument, that such adjustments to Coffman's model were unwarranted.<sup>20</sup> Thus, the Court will focus its analysis on August 9, 2001 and December 7, 2001, which both experts agree revealed a statistically significant price movement. Because the Court also finds lack of price impact on the other dates, for additional reasons, those dates are also discussed in further detail below.

### **3. Multiple Comparison Adjustment**

According to Allen, the multiple comparison issue arises when a large number of price reactions are tested for statistical significance, because the more price reactions tested, the greater the odds are of finding statistical significance simply due to chance. Allen Rep. at 13 ¶

23. Allen explained:

[I]magine rolling a 20-sided die with 19 white sides and 1 red side. If the die is rolled once, it would be surprising if the die landed red-side up since the likelihood of this occurring is only 5% (1 out of 20). However, if the die is rolled 100 times, it would be

---

<sup>20</sup> In his report, Coffman argues that Allen adjusts for too many events, and the only relevant events to test are the alleged corrective disclosures, of which there are six, not thirty-five. Coffman Rep. at 30 ¶ 56. Even if the Court were to accept this argument as valid, it does not address Allen's consistency argument—if Coffman was going to only test six dates, he should have included all untested dates in his control group.

much less surprising that the die landed red-side up 1 time since, on average, the die should land red-side up 5 times for every 100 rolls. Allen Rep. at 14 ¶ 24.

Allen argues that the multiple comparison problem arose in this case because the Fund asserted thirty-five dates on which alleged fraud inflated the stock price due to a misrepresentation or deflated the stock price because of a corrective disclosure. *Id.* at 14 ¶ 25. She calculated that each of thirty-five price reactions had a 5% chance of being found statistically significant because of normal fluctuations in the stock price. *Id.* Allen explains that, with that volume, it would not be surprising to find a statistically significant result on a day when no company-specific news was released to the market, and that the problem was exacerbated in this case by Nettesheim's initial method of testing all dates during the class period for statistical significance before looking for company-specific news.<sup>21</sup> *Id.* at 14, n.28.

Allen claims a commonly accepted method to correct for the multiple comparison problem is to apply the "Bonferroni adjustment."<sup>22</sup> *Id.* at 15 ¶ 26 (citing Hervé Abdi, *ENCYCLOPEDIA OF MEASUREMENT AND STATISTICS* 103–107 (Neil J. Salkind ed. 2007)). The Bonferroni adjustment takes into account the number of tests performed and adjusts the statistical thresholds of individual tests accordingly. *Id.*

---

<sup>21</sup> At the evidentiary hearing, Allen explained that if you test thirty-five dates at the 5% level, there is an 83% chance that you will find a statistically significant date due to chance. Hr'g Tr. at 34:10-12. However, if you test all dates in the class period, the chance of finding statistically significant price movement at the 5% confidence level increases to 99.99%. *Id.* at 34:22-25. In short, the more dates you test, the easier it is to find something statistically significant at the 5% confidence level. *Id.* at 35:3-5.

<sup>22</sup> Allen alternatively applied the so-called Sidak adjustment, which she found did not change her conclusions. Allen Rep. at 15 n. 29; Hr'g Tr. at 33:22-25. In other words, she found the statistical significance result to be the same under either the Bonferroni or Sidak adjustment. Hr'g Tr. at 148:22-25. The Sidak adjustment is considered to be very similar to the Bonferroni adjustment. Abdi, *supra* at p. 23, 108 (explaining that the Bonferroni adjustment is more well-known and more heavily cited than the Sidak adjustment).

Coffman argues that Allen's use of a multiple comparison adjustment is novel, improper, and yields erroneous results, because it results in unacceptably high false negatives.<sup>23</sup> Coffman Rep. at 4 ¶ 9. He claims that Allen's method requires stock price movement to be statistically significant at greater than 99% confidence instead of the generally accepted 95% threshold.<sup>24</sup> *Id.*

Coffman contends that the standard approach among economists for determining the statistical significance of an event is to evaluate the probability of false positives for that particular event, given the observed price movement, and then compare it to the generally accepted 95% confidence threshold. Coffman Rep. at 25 ¶ 45. Coffman states that he has never seen a multiple comparison adjustment, like that applied by Allen, used in a securities case, with the exception of one case in which one of Allen's colleagues was a designated expert. *Id.* at 26 ¶ 46. Coffman responds that the multiple comparison problem is not present in this case, because there is no risk of "data mining," which he concludes arises when data is randomly tested.<sup>25</sup> *Id.* at 26 ¶ 47. Rather, he argues there is a clear theory in this case and objective criteria for determining which data to test—in a semi-strong efficient market like that present here, there is an expectation that stock prices will react negatively to the revelation of negative news. *Id.* at 26–27 ¶ 48; Hr'g Tr. at 171:10-14. He argues that on every one of the six dates he tested, there

---

<sup>23</sup> Coffman also refers to this as a "Type II" error. Type II errors refer to the probability of saying there was not a price reaction when, in fact, there was a price reaction. Coffman Rep. at 4 ¶ 9.

<sup>24</sup> Coffman claims that Allen's application of the Bonferroni adjustment requires the price reaction to be significant at the 99.86% level, and only 6 out of 633 days during the class period would qualify as statistically significant under Allen's model. Coffman Rep. at 27–28 ¶ 52; Hr'g Tr. at 174:4-8. He notes that Allen herself testified that one would expect to observe 5% of the days to be statistically significant on chance alone. *Id.*

<sup>25</sup> Although Coffman uses the term "data mining," he did not elaborate on the concept or cite literature to explain the concept in the context of event studies. As far as the Court can tell, data mining is a very general term that is used in a wide variety of contexts in business and statistics. Moreover, none of the literature cited by Allen uses the term "data mining."

was news about Halliburton's asbestos exposure or accounting information, and therefore, a multiple comparison adjustment is unnecessary. Hr'g Tr. at 171:15-19.

Alternatively, Coffman argues that if a multiple comparison adjustment is used, the appropriate adjustment is a Holm-Bonferroni adjustment, because it lessens the high probability of false negatives that arises with the more conservative Bonferroni adjustment. Coffman Rep. at 29 ¶ 53; Hr'g Tr. at 172:10-13.

The Court is persuaded that the use of a multiple comparison adjustment is proper in this case because of the substantial number of comparisons, thirty-five comparisons, being tested for statistical significance in Allen's analysis. *See* Charles Seife, "The Mid-Reading Salmon," *Scientific American*, Aug. 2011, 30 (explaining that, in instances where as few as 20-40 comparisons are made, researchers are virtually guaranteed to find statistical significance in results that are, in fact, "statistical flukes"). Moreover, there is the unverified, but not entirely refuted, specter that Mr. Coffman's predecessor, Ms. Nettesheim, selected her dates by looking for statistically significant dates and *then* looking for Halliburton-specific news on those dates, from which Mr. Coffman selected the six events in his expert report. Coffman argues that the dates analyzed were specifically chosen because of news related to the allegations, not because of statistical price movement. However, that argument does not refute Ms. Allen's point. She concedes that the six dates were chosen because of news related to the allegations, but she argues they were selected *after* the thirty-five dates were found to have statistical price movement. Coffman did not explain how Nettesheim chose her dates, nor did he go into detail about how he chose the six dates.

However, the Court is mindful of Coffman's contention that the Bonferroni adjustment is overly concerned with Type I errors—false positives—and thus generates a relatively high



incidence of Type II errors—false negatives. *See* Coffman Rep. at 30 ¶ 55; *see also* Abdi, *supra* p. 23, at 111 (explaining that the Bonferroni adjustment becomes “very conservative when the number of comparisons becomes large and when the tests are not independent”). Thus, to the extent the measure of price impact is affected by the use of the Bonferroni adjustment rather than the Holm-Bonferroni adjustment, the Court will apply the Holm-Bonferroni adjustment, because it addresses the multiple comparison problem raised by Allen, while also guarding against the prospect of unacceptably high levels of Type II errors warned about by Coffman. Although the Court is mindful of the Fund’s argument that multiple comparison adjustments are rarely utilized in event studies for securities litigation, that argument does not refute the plausible concerns raised by Allen. The Court also notes that on certain of the dates in question, the absence of price impact is shown without making a Holm-Bonferroni adjustment. In conclusion, the Court finds that applying a Holm-Bonferroni multiple comparison adjustment is appropriate in this case.

#### **4. Additional Index**

To adjust for movement in Halliburton’s industries, Allen selected an index for each of Halliburton’s two main lines of business—(1) energy services, and (2) engineering and construction (E&C). To control for the energy services industry, she used the S&P 500 Energy Index, which is an off-the-shelf index used by Halliburton in its SEC filings to gauge its relative stock performance (“S&P Energy Index”). Allen Rep. at 12 ¶ 20. To control for the E&C industry, Allen used an index composed of Fortune 1000 companies classified by Fortune as being in the E&C industry (“Fortune E&C Index”). *Id.* Allen explained that she considered constructing an E&C index using analyst reports and financial news issued before and during the class period, but found the Fortune E&C Index had the “best fit” during the class period. *Id.* at

12 n. 20. Allen also tested whether Halliburton's stock price movement during the class period could be better explained by an index constructed of companies discussed by analysts as having asbestos exposure; however, she found that such an index did not add any more explanatory power to her model than did the S&P Energy Index and the Fortune E&C Index. *Id.* at 12 n.21; *id.* at Ex. 2.

Coffman argues that Allen's model does not adequately control for stock price movements in Halliburton's specific industries, and so her model suffers from what he calls "omitted variable bias."<sup>26</sup> Instead of merely using the S&P Energy Index and the Fortune E&C Index, Coffman argues Allen should have included a peer index composed of companies identified by securities analysts as being Halliburton's peers. Coffman Rep. at 3–4 ¶ 9. Coffman claims to have dramatically increased the explanatory power of Allen's event study by including such an index. *Id.* at 4 ¶ 9. Coffman describes the S&P Energy Index as being primarily comprised of oil companies. For example, the top three companies in the S&P Energy Index are oil-refining companies, not energy-services companies like Halliburton. Coffman Rep. at 17 ¶ 28. Therefore, an event that causes Halliburton's stock price to react might not similarly affect an oil company, and vice versa. *Id.* at 17–18 ¶ 29.

As a result, Coffman reviewed analyst reports and constructed a peer index composed of companies identified by analysts as being Halliburton's peers ("Analyst Index").<sup>27</sup> *Id.* at 18 ¶ 30;

---

<sup>26</sup> Omitted variable bias is error that occurs from failing to control for an important determinant of the variable of interest, which often results in drawing inappropriate statistical conclusions. Coffman Rep. at 3 n. 5. In other words, Coffman argues that, because Allen's study lacks the additional Analyst Index as an independent variable, the dependent variable, *i.e.* price reaction, is unreliable.

<sup>27</sup> Allen confirmed that this was a valid approach for constructing an industry index. Coffman Rep. at 18 ¶ 30. However, Allen stated that using such an index did not improve the explanatory power of her model. Allen Rep. at 12 n.20-21. In her rebuttal, Allen ultimately applied the

*id.* at Ex. 2. According to Coffman, analysts covering Halliburton discussed Baker Hughes and Schlumberger far more than any other companies.<sup>28</sup> *Id.* at 18–19 ¶ 31. Coffman found that those two companies only represented 6.2% of the S&P Energy Index utilized by Allen. Coffman Rep. at 19 ¶ 32, Ex. 3a.

On the other hand, the Analyst Index is a value-weighted index comprised of companies cited by analysts as Halliburton’s peers at least three times during the class period, and having a market capitalization of at least \$1 billion.<sup>29</sup> *Id.* at 19 ¶ 33. Baker Hughes and Schlumberger represent over half of the Analyst Index. *Id.* at Ex. 3a. Coffman argues, and provides evidence, that the addition of the Analyst Index better explains Halliburton’s stock price movement.<sup>30</sup> *Id.*

---

Analyst Index while making the internal consistency adjustments to Coffman’s model. Hr’g Tr. at 48:17-21.

<sup>28</sup> The Court takes judicial notice of the fact that Halliburton and Baker Hughes have recently merged. *See* March 27, 2015 Halliburton Press Release, “Halliburton and Baker Hughes announce approval of transaction by stockholders of both companies,”

<http://www.halliburton.com/en-US/news/hal-acquisition.page?node-id=hfc1272b>. *See* Fed. R. Evid. 201(b)-(c) (permitting a court to take judicial notice of an undisputed fact established by unquestionable sources or that is generally known within the court’s territorial jurisdiction).

<sup>29</sup> Coffman did not explain exactly where the Halliburton citations were located, either in analyst reports or elsewhere. The Court will assume that the citations to which Coffman refers were in analyst reports.

<sup>30</sup> Coffman stated that the adjusted R-squared statistic, which increases if a new term improves a model more than would be expected by chance increases from 49% to 74%, which the Court interprets to mean Allen’s model explains only 49% of the variance in Halliburton’s stock price during the class period, while adding the Analyst Index explains 74% of the variance. Coffman Rep. at 20–21 ¶¶ 34-35. Coffman also states that adding the Analyst Index also reduces the Root Mean Squared Error (RMSE) from .0228 to .0162, which the Court interprets to mean there is less unexplained “randomness” in the model with the Analyst Index. *Id.* at 21 ¶ 36. In contrast, Coffman contends that Allen’s regression model attributed 40% more of Halliburton’s stock price volatility to “randomness,” rather than to industry effects. *Id.* Coffman also found that the “coefficient,” which measures the magnitude of the influence that the particular index has on Halliburton’s stock price, is higher for the Analyst Index than it is for the S&P Energy Index or the Fortune E&C Index. *Id.* at 21–22 ¶ 37. Finally, he found that the t-statistic, which is used to measure whether the relationship being measured is unlikely to occur by chance, indicates Halliburton’s correlation with the Analyst Index is statistically significant at greater than a 99.99% confidence level. *Id.* at 22 ¶ 37.

at 20 ¶ 34. Allen did not meaningfully rebut that evidence.<sup>31</sup>

Because the addition of the Analyst Index constructed by Coffman increases the explanatory power of Allen's model, the Court is persuaded that it should be utilized in measuring the statistical significance of the price reaction on the six dates in question.

### 5. One or Two-Day Windows

On several of the six dates chosen by Coffman to measure price impact, he used a two-day window to measure price impact. For example, Coffman argues that a price decline on December 22, 2000 is related to a corrective disclosure on December 21, 2000. He argues that use of a two-day window is widely observed in financial literature, and Allen herself used a two-day window when comparing Halliburton's returns to other asbestos companies in the wake of the December 7, 2001 corrective disclosure. Coffman Rep. at 10–11 ¶ 13. Nevertheless, Coffman argues that all but one event has a statistically significant return, with greater than 95% confidence, over a one-day window, with December 21, 2000 being the only exception.

The Court finds that, in this case, the use of a two-day window is inappropriate to measure price impact in an efficient market. An efficient market is said to digest or impound news into the stock price in a matter of minutes; therefore, an alleged corrective disclosure released to the market at the start of Day 1, coupled with an absence of price impact throughout

---

<sup>31</sup> Hr'g Tr. at 151:12-15 (Q: "Would the explanatory power of your [re]gression analyses be increased by adding to your model to the Coffman peer group?" A [Allen]: "Yes. Mr. Coffman says the explanatory power increases and it does increase when he adds an additional index."); *id.* at 152:14-18 (Q: "And does the RMSE change in your model if you use Mr. Coffman's index?" A [Allen]: "Yes. The error goes down, which is what makes more things appear to be statistically significant than they would be if you didn't include the index."). Allen argued in part that the "more [Coffman] puts stuff in to [his model] to explain Halliburton's stock price, the more any deviation from the stock price then becomes statistically significant," which she says is akin to data mining. *Id.* at 151:17-21. She also noted that, in Coffman's prior reports, he used off-the-shelf indices and looked at what companies compared themselves to in SEC filings, as Allen did here. *Id.* at 48:17-21.

Day 1, followed by a price impact on Day 2, will not show price impact as to the alleged corrective disclosure. *See* Allen Rep. at 59 n.135 (citing Richard A. Brealey & Stewart C. Myers, *Principles of Corporate Finance*, at 351–53 (McGraw-Hill: New York, 7th ed. 2003)) (explaining that studies show that when firms publish their latest earnings or announce dividend changes, the major part of the price adjustment occurs within 5 to 10 minutes of the announcement); *see also* Fox, *supra* p. 8, at 444 n. 20 (“[W]e can assume that the predictive value of any firm-specific information that becomes newly public is reflected in price very quickly.”). As is discussed below, this principle negates the Fund’s allegation of price impact on December 21, 2000 and Halliburton’s argument that there was no price reaction on the second trading day after Halliburton’s December 7, 2001 disclosure.

## **6. Events at Issue**

### **A. Fixed-Price Construction Contracts**

On December 21, 2000, Halliburton issued a press release announcing that it would take a \$120 million after-tax charge because of restructuring and charges on projects in its engineering and construction business. Def. App. 336–37 (12/21/00 Press Release). According to the Fund, Halliburton’s 1999 10-K, released on March 14, 2000, disclosed “[c]laims and charge orders which are in the process of being negotiated with customers, for extra work or changes in the scope of work are included in revenue when collection is deemed probable.” Allen Rep. at 46 ¶ 94. The Fund alleges that this disclosure was a misrepresentation because Halliburton allegedly included claims in revenues even when their collection was not probable. *Id.* The truth was allegedly revealed to the market in a series of corrective disclosures, one of which was Halliburton’s December 21, 2000 Press Release announcing a \$120 million after-tax charge.

Allen argues there was no statistically significant price reaction on this date, both with and without adjusting for multiple comparisons. Allen Rep. at 47 ¶ 95. She also argues the disclosure was not corrective, and the market did not see it as such.<sup>32</sup> *Id.* at 49 ¶ 97; *id.* at 55–56 ¶¶ 116–21. Furthermore, Allen faults Coffman for using a two-day window to measure price reaction, arguing that Coffman’s approach ignores market efficiency, on which the *Basic* presumption relies. *Id.* at 59 n.135; Hr’g Tr. at 54:23-25. Allen argues that the press release was sent out at 8:58 a.m., *before* the market opened at 9:30 a.m., on December 21, 2000, and Coffman could only show price impact during the end of the day on December 22, 2000, nearly two days after the information entered the market. *Id.* at 58–59 ¶¶ 128–29; Hr’g Tr. at 50:23–51:1. Even still, after making a multiple comparison adjustment, Allen finds there was not a statistically significant price impact on December 22, 2000.<sup>33</sup> *Id.* at 49 ¶ 98.

Coffman found that, using a two-day window, with no multiple comparison adjustment, there was a negative and statistically significant decline in Halliburton’s stock price following the December 21, 2000 press release. Coffman Rep. at 10 ¶ 13; *id.* at 83 ¶ 182. Coffman also noted that Allen used a two-day window to show a lack of price impact on December 7, 2001. Coffman Rep. at 84 ¶ 184. Further, Coffman found statistically significant intraday movement during the day on December 21, 2000, at a 90% confidence level. *Id.* ¶¶ 185–86. Coffman agreed that there was no statistically significant stock price reaction on December 21, 2000, and

---

<sup>32</sup> However, as already discussed, the Court does not address this argument because it is more properly addressed at the pleading stage or merits stage of litigation, not the class certification stage.

<sup>33</sup> To the extent the Fund argues that Halliburton’s financial statements released on January 30, 2001 increased the December 21, 2000 charge from \$120 million to \$193 million, Allen contends that the \$193 million was pre-tax, and therefore consistent with the December 21, 2000 release. Allen Rep. at 60 ¶¶ 132–33. She notes that there was no price reaction on January 30, 2001, with or without adjusting for multiple comparisons. *Id.*; Allen Rep. at Ex. 1.

that his event studies generally use closing prices rather than intraday prices. *Id.* at 83 ¶ 183; Hr’g Tr. at 185:1-10.

Having considered the evidence provided by Allen and Coffman, the Court finds that there was no price impact on Halliburton’s stock following the December 21, 2000 disclosure, assuming it was corrective. Absent a compelling explanation, which was not given, the Court finds that the use of a two-day window is inconsistent with an efficient market, especially where the relevant disclosure was made before the market opened on Day 1. Even without adjusting for multiple comparisons, Coffman found an intraday statistically significant price reaction on Day 1 only at a 90% confidence level, which is less than the 95% confidence level both experts require in their regression analyses and which the Court finds is necessary. *See* Hr’g Tr. at 58:14-18, 59:1, 201:16-202:8 (Coffman agreeing that the Reference Guide on Multiple Regression, published by the Judicial Conference of the United States, requires that the level of statistical significance be 95%). In contrast, with and without a multiple comparison adjustment, Allen found no price impact on December 21, 2000. The Court agrees with Halliburton that there was no price impact on December 21, 2000, and finds that Defendants have rebutted the *Basic* presumption as to the allegedly corrective disclosure made on that date.

## **B. Asbestos Disclosures**

The Fund performed event studies for two types of disclosures relating to Halliburton’s asbestos liability—(1) Halliburton’s disclosures relating to the liability of its subsidiary, Dresser, for asbestos claims incurred by Dresser’s former subsidiary, Harbison-Walker (“Harbison”), which allegedly revealed that Halliburton knew Harbison needed financial assistance, but declined to disclose the full extent of Harbison’s request for assistance; and (2) a series of



adverse asbestos verdicts and judgments against Dresser, which allegedly revealed that Halliburton knew its asbestos liability was more significant than it had previously disclosed.

On May 14, 1999, in its 10-Q, Halliburton reported that its subsidiary, Dresser, had potential asbestos liability for claims filed against its former subsidiary, Harbison. Def. App. 255 (5/14/99 10-Q at 7). In early 1999, Halliburton had disclosed that Harbison had commenced arbitration, disputing its responsibility to indemnify Dresser for these claims. *Id.* On August 13, 1999, in its next 10-Q, Halliburton stated that Harbison was claiming it was owed \$40 million for amounts it had previously paid to resolve post-1992 asbestos claims. Def. App. 261 (8/13/99 Halliburton 10-Q at 8).<sup>34</sup> On November 9, 2000, Halliburton reported that it and Harbison had settled, and that Harbison had agreed to continue defending post-1992 claims and acknowledged its obligation to indemnify Dresser. Def. App. 330–32 (11/9/00 Halliburton 10-Q at 10–11).

**a. June 28, 2001**

On June 28, 2001, Halliburton informed investors of a “new development,” that Harbison had approached Halliburton seeking claims management and financial assistance concerning the post-1992 claims. Def. App. 367–68 (6/28/01 Halliburton Press Release at 1–2). On July 25, 2001, Halliburton announced it would take over management of the post-1992 claims against Dresser, and that Halliburton would take a \$60 million after-tax reserve for those claims. Def. App. 370 (7/25/01 Halliburton Press Release at 2).

Halliburton argues that its June 28, 2001 disclosure of Harbison’s request for claims management and financial assistance did not correct any alleged misrepresentation, and thus

---

<sup>34</sup> Allen found no statistically significant price reaction to this announcement under her model or Coffman’s model. Hr’g Tr. at 67:5–9. Of course, there is an important distinction between disclosing the existence of a demand or dispute and disclosing the fact that the demand would, indeed, be satisfied by Halliburton, and the latter happened on June 28, 2001.

there can be no price impact on that date. Allen Rep. at 122 ¶ 281. Halliburton argues that investors were aware of the indemnity relationship between Harbison and Dresser throughout the class period, and Halliburton had already disclosed the risk that it would have to pay Harbison's post-1992 claims, with no price reaction. *Id.* at 123 ¶ 283 (citing 1999 10-Q, filed May 14, 1999); Allen Rep. at Ex. 1. Furthermore, Halliburton says Ms. Allen's opinion confirms as much, because she found no evidence that market analysts viewed the June 28, 2001 disclosure as indicating that Halliburton's prior representations had been misleading. Allen Rep. at 126 ¶ 290. Finally, Allen's study found no statistically significant price drop that day, after making a Bonferroni multiple comparison adjustment, and she found no statistically significant price drop under Coffman's model after making internal consistency adjustments. Allen Rep. at 125 ¶ 289; Hr'g Tr. at 63:19-25.

The Fund argues that the June 28, 2001 press release was corrective of Halliburton's fraudulent omission that it knew Harbison would be seeking financial assistance, and that it knew there would likely be an increase in the number of asbestos claims filed against it. Dkt. No. 594 at 19. According to the Fund, the June 28, 2001 press release indicated that Halliburton's costs for resolving at least some of its pending claims would rise, which made the release corrective of Halliburton's prior misrepresentations about its liability for pending claims. Coffman Rep. at 69–70 ¶¶ 147–49. Coffman notes that Allen cited to an analyst report from A.G. Edwards, in which the analyst connected the June 28, 2001 press release to Halliburton's representations in previous SEC filings. *Id.* at 70 ¶¶ 147–48 (quoting Allen Rep. at ¶ 287). Furthermore, Coffman claims that many of Allen's arguments focus on whether there was a material misrepresentation, when the inquiry should be to assume a material misrepresentation, and then determine whether there was price impact. Coffman Rep. at 72 ¶ 153. Despite Allen's

argument that analysts did not understand Halliburton's June 28, 2001 announcement as revealing a prior falsity, Coffman maintains it is not necessary for analysts to contemporaneously recognize the prior announcement as false, as Allen agreed. Coffman Rep. at 72–73 ¶ 154 (citing Allen Dep. at 103:5-10). Coffman found a negative and statistically significant price reaction in both the one and two day windows following the June 28, 2001 disclosure, and he claims that Allen did as well, before she applied the Bonferroni adjustment. Coffman Rep. at 70–71 ¶¶ 150–51.

Having considered the evidence provided by Allen and Coffman, the Court finds there was no price impact relating to the June 28, 2001 disclosure. As already stated, the Court will limit its inquiry to the event studies, not to the factual matter of whether the disclosure was corrective. The issue regarding this event is whether the Court should apply the Holm-Bonferroni adjustment to Allen's model and whether the "internal consistency" adjustments advocated by Allen are appropriately applied to Coffman's model. As already discussed, Allen made each of three adjustments advocated by Coffman—using the additional Analyst Index, making a multiple comparison adjustment for only six dates, and applying the Holm-Bonferroni adjustment—and added in the balance of the thirty-five dates to make Coffman's model internally consistent, and she found no price impact. Hr'g Tr. at 48:7-11. As already discussed, the Court finds that these adjustments are appropriately applied to Coffman's model. Accordingly, neither Coffman's, nor Allen's, analysis shows price impact on June 28, 2001, and Defendants have rebutted the *Basic* presumption as to the corrective disclosure on that date.

#### **b. August 9, 2001**

On August 9, 2001, Halliburton filed its second quarter 2001 10-Q, in which it explained that, during that quarter, Halliburton "experienced an upward trend in the rate of new asbestos

claims” being filed. Pl. App. 388 (Halliburton 2Q01 SEC Form 10-Q). Halliburton also explained that its gross asbestos liability had grown to \$699 million as of June 30, 2001, which was much larger than the \$60 million in liability Halliburton had announced on July 25, 2001.<sup>35</sup> Coffman Rep. at 74 ¶ 159 (citing Pl. App. 388). Coffman claims that Halliburton’s 10-Q disclosure provided additional detail regarding Halliburton’s asbestos exposure. *Id.* at 73 ¶ 157.

Halliburton argues that the 10-Q was actually confirmatory of information Halliburton had previously reported, and therefore not a corrective disclosure nor evidence of price impact. Halliburton argues that the \$699 million figure disclosed in the August 9th 10-Q appears much larger than the figure previously announced, but the figures are in fact the same because the previously-announced figure was net of insurance.<sup>36</sup> Hr’g Tr. at 73:22-23. On July 25, 2001, Halliburton disclosed the \$60 million figure as a net amount because market analysts stated that Halliburton had approximately 75% of its claims covered by insurance, and the Harbison claims were 90% covered. *Id.* at 74:7-11, 221:7-25. Therefore, on July 25, 2001, the market could have easily surmised that Halliburton’s gross asbestos liability was around \$699 million, and there

---

<sup>35</sup> The Fund initially claimed that the August 9, 2001 10-Q disclosed for the first time that Halliburton had increased its asbestos reserves from \$30 million to \$124 million, but Coffman now concedes that this information was already disclosed to the market on July 25, 2001. Coffman Rep. at 73–74 ¶ 158; Hr’g Tr. at 218:17. Halliburton’s net asbestos liability as of March 31, 2001 was \$ 30 million. Allen Rep. at 126–27 ¶ 293. The \$124 million in net liabilities reported on August 9, 2001 was an increase of nearly \$95 million from the previous period. *Id.* However, this increase was almost entirely due to the \$92 million pre-tax amount the company decided to record for pending Harbison claims, which it had already announced in a press release issued July 25, 2001. *Id.* The \$60 million in the August 9, 2001 10-Q was the exact same information announced in the July 25, 2001 press release. Allen Rep. at 127 ¶ 195; Hr’g Tr. at 71:8-11, 72:15-16.

<sup>36</sup> The \$60 million net liability previously disclosed approximates to \$92 million pretax, and applying the 90% coverage rate to the \$92 million pretax figure results in \$920 million gross liability, which does not include the \$84 million Halliburton had disclosed in non-Harbison related claims in the prior quarter. In any event, \$920 million is substantially larger than the \$699 million the Fund claims was a corrective disclosure. Hr’g Tr. at 221–22.

was no statistically significant price reaction on that date according to Coffman's and Allen's respective models. *Id.* at 79:1-4.

Halliburton also rejects the Fund's attempted focus on an alleged increase in the rate of new asbestos claims being filed, because it argues that the rate of new claims being filed increased at a relatively consistent rate in each of the previous quarters in 2001. Hr'g Tr. at 77:21-25. Halliburton argues that it never stated that its claims or rate of claims would never increase. *Id.*

After adjusting for multiple comparisons, Allen argues there was no statistically significant price reaction on August 9, 2001. *Id.* ¶ 299; Hr'g Tr. at 70:6-7. On the other hand, Coffman found a statistically significant price reaction before the application of a multiple comparison adjustment. Coffman Rep. at 74–75 ¶ 161.

The Court finds that, on this date, Halliburton has met its burden of showing a lack of price impact, if only because the Fund has not shown that Halliburton disclosed any information related to its asbestos liability that was not already impounded in the market price of the stock on August 9, 2001, and therefore, there can be no price reaction. Halliburton effectively rebutted the Fund's claims that Halliburton disclosed either a substantial increase in its gross asbestos liability or a substantial increase in the rate of new claims. Halliburton has demonstrated that the gross liability figures were akin to comparing apples and oranges, in that the previously disclosed figures were after-tax and/or net of insurance, and Halliburton's rate of insurance coverage was well known to the market. Coffman contends that his studies show that it was "highly unlikely that the market anticipated everything that was announced" on August 9, 2001, because his study shows a statistically significant price reaction at the 99.99% level on that date. Hr'g Tr. at 226:5-8. The Fund has shown that there was a price movement on that date;

however, Halliburton has demonstrated that the disclosure that allegedly caused the price reaction was already disclosed to the market on July 25th, to no price reaction. The Court's finding is reinforced by Coffman and the Fund's concession that the Fund's previous alleged corrective disclosure—that net asbestos liability increased from \$60 million to \$124 million—had already been disclosed on July 25, 2001. Coffman's new theories relating to August 9, 2001 fare no better under closer scrutiny. The Court is not determining as a matter of law that the disclosures were not corrective, but rather, that Halliburton has shown that the information alleged by the Fund to be corrective was *both* already disclosed *and* caused no statistically significant price reaction. Thus, the Court finds that Halliburton has rebutted the *Basic* presumption with respect to the corrective disclosure on August 9, 2001.

**c. October 30, 2001**

On October 26, 2001, a Mississippi jury found Dresser, AC&S Inc., and 3M Corp. liable for damages totaling \$150 million, one of the largest verdicts ever rendered in asbestos litigation at the time. Allen Rep. at 88 ¶ 202. Dresser's share was \$21.25 million, and Halliburton announced the verdict in a press release issued on October 30, 2001. *Id.*

Allen notes that the verdict was first announced on Sunday, October 28, 2001 in the *Clarion-Ledger*, a Mississippi statewide newspaper, and Associated Press wires. *Id.* at 88–89 ¶ 203, 207; Hr'g Tr. at 84:7-9. Allen argues that this disclosure on October 28, two days before Halliburton's press release of October 30, negates any price reaction on October 30 or 31, because an efficient market would have already absorbed the news of the verdict and impounded that news in the stock price *before* Halliburton's press release was issued. Hr'g Tr. at 82:24-25, 85:17-19. Allen found no statistically significant price reaction on October 29, when Dresser's codefendant in the case, 3M Corp., made a public announcement about the verdict that

mentioned Dresser. *Id.* at 89 ¶¶ 204-05. According to Allen, there is significant overlap between Halliburton and 3M investors—more than half of Halliburton’s outstanding shares are owned by investors that also own 3M shares—so one would assume that many investors that took notice of the 3M press release would also be interested in Halliburton. Hr’g Tr. at 86:16-20.

Allen also notes that analysts later commented on the lack of price reaction following the Mississippi verdict. Allen Rep. at 89 ¶ 206 (quoting Credit Suisse Report, 12/10/01) (“Looking at Halliburton’s share price performance over the past year it is interesting to note that it didn’t move – up or down . . . when the company made the filing for the second case, on the 30th of October.”). Despite the verdict’s unprecedented size, Allen argues that analysts did not change their outlook on Halliburton’s asbestos liability, and instead noted that the asbestos litigation environment was unpredictable and irrational. Allen Rep. at 90 ¶ 208 (quoting Salomon Smith Barney Report, 10/31/01) (“The disclosure . . . does not appear to signal a meaningful change in the pattern of asbestos litigation for the company . . . we recognize the unpredictability and apparent irrationality of the asbestos litigation environment in general.”).

Finally, Allen found no statistically significant price reaction following the October 30, 2001 announcement, after making a Bonferroni adjustment for multiple comparisons. Allen Rep. at 90 ¶ 210; *id.* at Ex. 1. Allen also found no statistically significant price reaction on October 31, 2001, the date the press release was issued, after she made internal consistency adjustments to Coffman’s model. Hr’g Tr. at 88:21-23.

Coffman argues that Allen’s event study shows a statistically significant decline in Halliburton’s stock price on October 31 and November 1, which coincides with Halliburton’s announcement of the Mississippi verdict. Coffman Rep. at 61 ¶ 127; Hr’g Tr. at 193:10-14 (Coffman explained that the verdict was announced by Halliburton after the close of the market



on October 30). Coffman claims Allen has provided no evidence that shows Halliburton's stock was reacting to some alternative information. *Id.* To the extent Allen argues there is no price impact on these days, Coffman attributes her findings to her utilization of a multiple comparison adjustment, which he finds flawed. Coffman Rep. at 62 ¶ 128.

Coffman tested October 31 and November 1, as opposed to October 28 and 29, because he argues that those were the first dates during that period that analysts and news outlets specifically talked about Halliburton's role in the Mississippi case. Hr'g Tr. at 193:20-23; Pl. Hr'g Ex. 19-20. Coffman argues that it is entirely plausible that the vast majority of market participants did not become aware of the verdict until Halliburton issued its own press release. Coffman Rep. at 62 ¶ 129. Coffman argues that many market participants may not monitor Mississippi newspapers, and it is unsurprising that those market participants might have missed the 3M announcement, because 3M is in a different industry, and the 3M announcement did not mention Halliburton by name. *Id.* Coffman argues that there were at least two analyst reports issued between the time of the article in the *Clarion-Ledger* and Halliburton's press release on the evening of October 30, neither of which mentioned the verdict. Hr'g Tr. at 193:24-194:7; Pl. Ex. 22 (Johnson Rice & Company L.L.C. Report, 10/29/01); Pl. Hr'g Ex. 23 (Deutsche Bank Alex. Brown Inc. Report, 10/29/01). In contrast, analyst reports on October 31 described the verdict as new information. Coffman Rep. at 62 ¶ 129, n.124; Pl. Hr'g Ex. 20 (Salomon Smith Barney Report, 12/31/01). Finally, Coffman argues that the two-day market capitalization decline of \$539 million far exceeded the financial impact of the verdict itself, \$21.25 million, which means the market was impounding additional information beyond the direct financial impact of the verdict. *Id.* at 63 ¶ 130. Coffman points out that Allen's asbestos index on those days showed

positive growth, which means that Halliburton's price decline cannot be attributed to general uncertainty about asbestos. *Id.*

The Court finds that Halliburton has met its burden and demonstrated a lack of price impact as to the announcement of the Mississippi verdict. The Court will not rule on whether the October 30 press release actually contained a corrective disclosure, for the reasons already discussed. However, the Court is persuaded that the absence of a price reaction on October 29, after the verdict had already been disclosed in a statewide newspaper, some AP wires, and 3M had disclosed the verdict in a press release which mentioned Dresser by name, all negates any finding of a price reaction on October 30 or 31. Public announcements preceded Halliburton's press release and, as already discussed, the Court is required to assume that the market had already absorbed that information by the time Halliburton made its own announcement on the evening of October 30. Mr. Coffman attempts to downplay the readership of the *Clarion-Ledger*, notes that the 3M announcement did not mention Halliburton by name, and argues that there was a dearth of analyst reports between the date of the *Clarion-Ledger* publication and Halliburton's own press release. Taken together, the presence of these disclosures and the absence of price impact on October 29 and 30 persuade the Court that Halliburton has showed an absence of price impact as to the October 30, 2001 press release regarding the Mississippi verdict.<sup>37</sup> Moreover, if Allen's internal consistency adjustments are applied to Coffman's

---

<sup>37</sup> The Court asked Ms. Allen whether there is a distinction between newspapers, and the speed in which some newspaper disclosures reach an efficient market. Hr'g Tr. at 85:20-87:5. Allen did not draw a firm distinction between a publication like the Wall Street Journal and a newspaper like the *Clarion-Ledger*. *Id.* However, she noted that a semi-strong market is supposed to assume that the market absorbs all public information, and here, there was a substantial overlap between 3M investors and Halliburton investors. *Id.* Mr. Coffman does not argue that the 3M announcement and associated news stories did not reach the market shortly after being released. As discussed, the 3M press release did, in fact, mention Dresser by name. *Id.* at 87:3-6.

analysis, along with the Holm-Bonferroni multiple comparison adjustment and the additional Analyst Index, there is no price impact on October 31 either, and the Court has already stated that it finds those adjustments to be appropriate. Thus, the Court finds that Halliburton has rebutted the *Basic* presumption with respect to the Mississippi verdict announced on October 30, 2001.

**d. December 4, 2001**

On December 4, 2001, Halliburton announced in an 8-K filing that on November 29, 2001, a Texas district court entered judgment against Dresser for its \$65 million share of a \$130 million verdict rendered on September 12, 2001 against Halliburton's subsidiary, Dresser, and another co-defendant. Allen Rep. at 86 ¶ 193, 91 ¶ 211. Halliburton also disclosed in the same 8-K filing that the same Texas district court entered three additional judgments against Dresser in favor of 100 other asbestos plaintiffs, awarding them \$35.7 million because of a breach of a settlement agreement signed earlier in the year by Harbison. *Id.* at 91 ¶ 211. The 8-K was released after the market closed on December 3 and before the opening bell on December 4. Hr'g Tr. at 176:5-6.

The earliest news story Allen could find about the September 12 verdict was a September 20, 2001 story in the *Beaumont Enterprise*, a daily paper covering Orange County, Texas, and she argues there was no statistically significant price reaction after that disclosure, despite the record size of the verdict and Halliburton's involvement. *Id.* at 86 ¶¶ 194-95. Allen also asserts that there was no price reaction to a story published on September 21, 2000 in Mealey's, which covers litigation and asbestos news. *Id.* at 86-87 ¶ 196. Her conclusions about the price reactions were based on her analysis before and after adjusting for multiple comparisons. *Id.* at 86-87 ¶ 195; *id.* at Ex. 1. Finally, she relied on analyst commentary months later that

mentioned the absence of price reaction after the September 12, 2001 verdict. *Id.* at 87 ¶ 197.

Thus, Allen argues that news reports and trade publications disclosed the verdict underlying the December 4 judgment months earlier, to no price reaction. Allen Rep. at 91–92 ¶¶ 214–15.

Allen also notes that analysts described the announcement on December 4th as widely known. *Id.* at 92 ¶ 217. As for the disclosure of judgments relating to Harbison’s breach of settlement agreements, Allen argues that the 8-K disclosure was not corrective, because Halliburton had made no prior representation about these cases, the market was well aware of the possibility of adverse rulings, and analysts did not find the disclosures to be corrective. *Id.* ¶¶ 218–19.

Allen found no statistically significant price reaction after the 8-K release on December 4, both before and after adjusting for multiple comparisons, and she also found no price reaction on December 5, 2001. Allen Rep. at 91–92 ¶¶ 211, 220; Hr’g Tr. at 88:16–18. Allen found no price impact based on Coffman’s model either, once she applied the internal consistency adjustments already discussed. Hr’g Tr. at 88:21–23.

In response, Coffman argues that there was in fact a stock price decline of nearly 5% on September 20, 2001, and after including his Analyst Index, there was a negative price reaction that is significant at the 90% confidence level. Coffman Rep. at 58 ¶ 119. Coffman notes that Halliburton’s stock fell by \$258 million, even though the verdict rendered against it was only \$65 million. *Id.* With respect to the December 4 disclosure, Coffman refutes Allen’s argument that the absence of a price reaction on September 20 negates a finding of price impact on December 4, and notes that the December 4 announcement and the price reaction to it were consistent with the market slowly learning more about Halliburton’s asbestos liabilities as time went on, with the December 4 announcement serving as the proverbial straw that broke the

camel's back. *Id.* at 60 ¶ 123. Coffman argues that Allen fails to explain why the \$35.7 million in judgments for Harbison's breach of settlement agreements was not new information or was unrelated to the Fund's claims, and the entry of judgment on the September 12 verdict signaled that the judge was not overturning the verdict. *Id.* at 64 ¶ 133. After adding the Analyst Index, Coffman found a statistically significant price decline over both one and two-day windows on December 4 and 5 at above the 99% confidence level. *Id.* at 63 ¶ 131.

The Court finds that Halliburton has met its burden of showing an absence of price impact as to the disclosures in the December 4 press release. First, the Court does not view the absence of price impact as to the September 12 jury verdict as dispositive, or even meaningfully relevant, if only because, as the Court explained during the hearing, there is an important distinction between the rendering of a verdict and the entry of a judgment. Moreover, the December 4 press release also disclosed that Halliburton would be liable for \$35.7 million in judgments relating to Harbison's breach of settlement agreements, and there was no corresponding news relating to those judgments disclosed on September 20. Thus, the Court will look only at whether there was a statistically significant price reaction on December 4, 2001. If Allen's internal consistency adjustments are applied to Coffman's model, there was no statistically significant price reaction on December 4. Hr'g Tr. at 88:21-23. The Court has already explained that these adjustments are appropriate; accordingly, the Court finds a lack of price impact on December 4, 2001 and Halliburton has met its burden of rebutting the *Basic* presumption with respect to the corrective disclosure made on that date.

**e. December 7, 2001**

On December 7, 2001, before the market opened, Halliburton issued a press release announcing that a Baltimore jury had returned a verdict in favor of five plaintiffs against three defendants, and Dresser's share of responsibility for the verdict was \$30 million. Def. App. 424 (12/7/01 Halliburton Press Release at 2); Allen Rep. at 93 ¶ 221. On that day, Halliburton's stock price dropped by approximately 40%, a decline that was statistically significant under both Coffman and Allen's models.<sup>38</sup> Hr'g Tr. at 91:6-10. However, Halliburton argues that the question is whether there was price impact *from the alleged misrepresentation*, not simply whether the price dropped from the announcement. *Id.* at 91:25-92:2. Accordingly, Allen concludes that there was no price impact as to the alleged misrepresentation. *Id.* at 91:19-20. Allen supports her conclusion by arguing that (1) Halliburton's stock rebounded on the next trading day, Monday, December 10, 2001, (2) the announcement did not include any new information regarding any of the alleged misrepresentations concerning Halliburton's asbestos liability, (3) analysts continued to believe Halliburton was effectively managing its asbestos liability and did not believe they had been previously misled, and (4) the price decline is more properly attributed to an increase in uncertainty, changes in the asbestos environment, and a spike in implied volatility that also affected other asbestos companies at the end of the class period. Coffman disputes each of the foregoing, and notes that Allen admitted in her deposition that she cannot attribute Halliburton's entire price decline to other factors, such as implied volatility. Coffman Rep. at 32–33 ¶ 61. Coffman argues that newly disclosed information—the

---

<sup>38</sup> Coffman explains that the “P value” for December 7, which is the probability of a false positive, is virtually zero and indicates that there is well over a 99.9 percent confidence that the stock price moved in response to the news announced on that date. Hr'g Tr. at 182:14-17.

Baltimore verdict—corrected Halliburton’s previous disclosures relating to its asbestos liability, and caused the price decline on December 7. *Id.*

The Court finds that Halliburton cannot show an absence of price impact based on the December 10 price rebound, because, as already discussed, the Court will not find an absence of price impact based on the returns during Day 2 of a disclosure made on Day 1, because to do so would be inconsistent with an efficient market, which is said to digest or impound news into the stock price in a matter of minutes. *See Fox, supra* p. 8, at 444 n. 20. Thus, Halliburton cannot show an absence of price impact by a rebound in the stock price on December 10. Furthermore, the Court will not determine as a matter of law whether the verdict announcement was corrective.<sup>39</sup> Thus, the Court will focus on the analyst reports addressing the December 7 announcement and Halliburton’s argument that general uncertainty in the asbestos environment caused Halliburton’s statistically significant price decline.

Allen argues that none of the analysts covering Halliburton indicated after the December 7 announcement that they had been misled, but instead, the analysts continued to think Halliburton was effectively managing its asbestos liabilities. Allen Rep. at 75 ¶ 167. Analysts believed Halliburton was keeping adequate reserves, insurance would cover the verdicts, and the verdicts would be overturned or modified because Halliburton did not become involved in the

---

<sup>39</sup> Halliburton argues that the Fund’s theory—that the disclosure caused the market to realize Halliburton had been misleading it—is contradicted by the reaction, or lack thereof, to previous, larger verdicts and judgments. Allen Rep. at 94 ¶ 224. The Texas and Mississippi verdicts were larger in total and Halliburton’s relative responsibility was larger, and there was no price reaction on September 20 or December 4, as already discussed. *Id.* Moreover, the Baltimore verdict disclosed on December 7 was for mesothelioma, which is always associated with greater damages awards because it is fatal and caused only by asbestos. *Id.* ¶ 225. As already discussed, the Court assumes at this class certification stage that Halliburton’s disclosure was, in fact, corrective. Moreover, unlike August 9 and October 30, where information had already been disclosed to the market, to no price impact, the Baltimore verdict had not been previously disclosed.



litigation until later in the process. Allen Rep. at 95 ¶ 226. On the other hand, Coffman cites to analyst reports that discussed the Baltimore verdict as new information. Coffman Rep. at 66 ¶ 138 (PNC Report 12/7/01) (“[W]ith recent high levels of judgments against the company . . . we are concerned that [Halliburton’s reserves] may prove to be low.”); *id.* at 67 ¶ 139 (Hibernia Southcoast Capital Report 12/7/01) (“Halliburton’s stock is trading down significantly today, as the Company announced yet another significant award against the Company for asbestos related litigation.”) (JP Morgan Report, 12/7/01) (“Shares of Halliburton are down sharply . . . on the sizeable award. This is the fourth significant judgment against Halliburton since late October. The size of this claim . . . is materially higher than Halliburton’s historical average cost per claim of less than \$200) (ABN-AMRO Report, 12/10/01) (“Halliburton’s shares plunged . . . as new negative news regarding the company’s asbestos problems poured into the market”) (Pl. Hr’g Ex. 18, Jeffries Report, 12/7/01) (“These are surprising developments . . . [w]e now believe that HAL’S net asbestos-related liabilities could be significantly higher than currently estimated . . . .”).

Coffman also relies on an online blog, TheStreet ([www.thestreet.com](http://www.thestreet.com)), which is a widely read blog run by Jim Cramer, who regularly provides investment advice on CNBC and other news outlets. Pl. Hr’g Ex. 11; Hr’g Tr. at 128:6-16. TheStreet post, titled “Halliburton Buried as Investors Stop Believing,” stated that “shares dove to nine-year lows Friday [December 7, 2001] as investors lost faith in the company’s claims that asbestos litigation would never catch up with it.” Pl. Hr’g Ex. 11; TheStreet, “Halliburton Buried as Investors Stop Believing,” <http://www.thestreet.com/story/10005091/1/halliburton-buried-as-investors-stop-believing.html> (last visited May 2, 2015). Allen argues that Halliburton never claimed that its asbestos litigation liabilities would never catch up with it, and TheStreet post is not an analyst report and

contains no attribution or analysis to support its quoted statements, and it is inconsistent with other analyst reports and commentary. Allen Rep. at 76 ¶¶ 169-71.

The analyst reports cited by Coffman show that the TheStreet post was not an aberration. Although some analyst reports on December 7 and 10 may have been more optimistic than others, the reports cited by Coffman and the Fund make it clear that a sufficient number of analysts viewed Halliburton's disclosure of the Baltimore verdict on December 7 as new news, and the cause of Halliburton's price decline. Thus, the analyst reports and commentary cited by Allen will not serve to refute what the parties agree was a statistically significant price reaction on December 7.

Allen argues that Halliburton's stock decline was attributed to an increase in uncertainty and change in the economic and asbestos environment that also affected other companies. Allen Rep. at 97 ¶ 229; Hr'g Tr. at 96:10-11 (noting that CBS Viacom's stock price also decreased). She notes that the stock price declines of other companies, at least in terms of market capitalization, were even larger than Halliburton's stock price decline, and market commentators discussed the price declines of Halliburton and CBS Viacom together. Allen Rep. at 102-03 ¶¶ 238-41. Allen argues there was uncertainty and volatility at the end of the class period, and when risk or uncertainty rises, stock prices fall. *Id.* at 98 ¶ 231. She notes that analysts lowered their expectations about Halliburton's stock price because of this increased uncertainty and changing market conditions. *Id.* at 99 ¶ 233; Hr'g Tr. at 96:1-3 (citing various analyst reports).

Allen claims there was a huge spike in Halliburton's implied volatility at the end of the class period, after the December 7 disclosure, yet there was no similar spike after Halliburton's disclosure of the Texas verdict or judgment or the Mississippi verdict. *Id.* at 99 ¶¶ 234-36.

Implied volatility is a measure of stock price volatility.<sup>40</sup> Allen also notes that companies with more pending claims had a larger drop in market capitalization and greater implied volatility following the December 7th announcement of the verdict.<sup>41</sup> Allen Rep. at 104–05 ¶¶ 244-45; Hr’g Tr. at 97:17-20, 103:11-14. Allen notes that Honeywell and 3M were co-defendants in the Texas and Mississippi cases that resulted in substantial verdicts and judgments, but neither experienced significant price reactions attributable to the verdicts and judgments during the class period; however, they did experience price reactions at the end of the class period. Allen Rep. at 107 ¶¶ 246-47. Dow Chemical had no disclosures, news reports, or analyst reports mentioning its asbestos liability during the class period, yet at the end of the class period through 2002, there were hundreds of news articles and analyst reports discussing its asbestos exposure. *Id.* at 112 ¶¶ 256-58. Dow Chemical nonetheless experienced a statistically significant price decline and an increase in its implied volatility at the end of the class period. *Id.* Allen argues that the price declines of each of these companies show that the stock prices of Halliburton and other

---

<sup>40</sup> Allen explained that option prices today tell investors what the market is thinking about the variability of a stock’s future price. Hr’g Tr. at 97:4-7. Implied volatility is just a measure of the market’s expectation of the future volatility of a stock price at a given point in time. *Id.* at 97:7-13. For example, if an option price rises, that means there is a high demand for a fixed future price, and a concomitant worry that the price could be volatile in the future. The value of an option is bigger if there is more volatility, i.e. it is “in the money.”

<sup>41</sup> Allen looked at an index of asbestos companies that she composed based on analyst commentary. She hypothesized that companies with more asbestos exposure would have greater reactions. Hr’g Tr. at 100:11-17. She compared the number of pending claims with the drop in market capitalization, and found that companies with more pending claims saw a larger drop in market capitalization. *Id.* at 102:6-8, 103:11-14. She argues that if the verdict was just a correction of Halliburton’s prior misrepresentation, it would not make sense for these other companies also to have significant stock price declines. *Id.* at 102:12-14. However, Allen fails to address why the disclosure of the verdict could not correct an earlier misrepresentation by Halliburton, thereby impacting Halliburton’s stock price, *and* adverse asbestos verdicts could also cause other asbestos companies to experience stock declines.

companies with asbestos exposure moved more closely together in 2002 than they did in 2001.<sup>42</sup> *Id.* at 116 ¶ 265. Thus, Allen argues that Halliburton's price decline was caused by an overall change at the end of the class period with respect to the effect of asbestos on stock prices.<sup>43</sup> *Id.* at 118 ¶ 271.

Coffman faults Allen for using raw dollar market capitalization to compare Halliburton to other companies over the December 7-10, 2001 period. He claims Allen's analysis is inconsistent with the rest of Allen's event study, which analyzes percentage changes, and Halliburton's relative decline is far larger than any of the other asbestos companies Allen identified. Coffman Rep. at 5-6 ¶ 10, 34-35 ¶ 65; Pl. Ex. 7-8a, Dkt. No. 591, Pl. App. 104-05. For example, CBS Viacom had a market capitalization of \$85 billion, and regularly experienced changes in market capitalization of \$2 billion, while Halliburton's market capitalization was only \$9 billion. Coffman Rep. at 35-36 ¶ 68.

Next, Coffman argues that Allen did not adequately test her assertion that observed price changes in other companies explain a substantial portion of Halliburton's price decline, and that Coffman's analysis shows that Allen's asbestos index of 31 companies explains less than 4% of Halliburton's stock price movement on December 7. Coffman Rep. at 6 ¶ 10, 39-40 ¶¶ 75, 77;

---

<sup>42</sup> Allen observed the relationship of 31 companies identified by analysts as having asbestos exposure in 2001 and 2002. Allen Rep. at 116 ¶ 265. She found that their relationship went from 25% during the class period to 61% after the class period, and all companies with asbestos exposure moved more tightly together after the class period. *Id.* at 116-17 ¶¶ 267-70; Hr'g Tr. at 104:4-5.

<sup>43</sup> Allen also asserts that the December 2, 2001 Enron bankruptcy contributed to Halliburton's price decline because it increased uncertainty at the end of the class period. Allen Rep. at 101 ¶ 237. Yet, as Coffman notes, Allen does not formally evaluate her assertion, and instead, only cites to two news reports and a single analyst report. Coffman Rep. at 44-45 ¶ 88-90. Thus, the Court is without the necessary statistical data to tie the bankruptcy to Halliburton's stock price decline. Regardless, it is unlikely that Enron's bankruptcy, which occurred five days before the December 7 announcement, would meaningfully impact the Court's analysis.

Pl. Ex. 11, Dkt. No. 591, Pl. App. 110. Coffman argues that Halliburton's stock price should have only declined 2.1% as a result of the change in the asbestos environment, given that its peers declined 2.3%, but Halliburton's stock price actually dropped 57% after controlling for Allen's other industry indices. *Id.* at 40 ¶¶ 76-77. However, Allen counters that you would not expect all of the companies to drop the same percentage amount, which is what Coffman's analysis assumes. Hr'g Tr. at 104:9-23.

Coffman also argues that greater implied volatility in Halliburton's stock price after the December 7 disclosure is a reflection of the impact of the corrective information itself, and not the cause of Halliburton's price decline, as Allen argues. Coffman Rep. at 7 ¶ 10. In other words, the corrective information created the uncertainty, and the analysts cited by Allen attributed the increased uncertainty to Halliburton's asbestos liabilities. Furthermore, Coffman argues that not all the companies in Allen's asbestos index experienced as dramatic an increase in implied volatility as Halliburton did, and Allen only focuses on a few companies to prove a point that is partially refuted by other companies in her own asbestos index. *Id.* at 43-44 ¶ 86; Pl. Ex. 13, Dkt. No. 591, Pl. App. 112. He contends that, even with respect to the companies Allen focuses on, such as Pfizer, she overstates the increase in implied volatility, which happened later and over a longer time period.<sup>44</sup> *Id.* at 43 ¶¶ 84-85. Allen counters that one would not expect the news to affect each company equally, and there is no reason to believe that the stock of all of the companies in her asbestos index would have a similar reaction. Hr'g Tr. at 100:1-10.

---

<sup>44</sup> For example, Allen used a different number of days in measuring the implied volatility of different companies in her asbestos index. Hr'g Tr. at 136:3-140:24. She claims to have done so because analysts discussed the different companies over different periods of time during this alleged period of uncertainty; however, her method casts further doubt on whether an increase in implied volatility negates the substantial price decline Halliburton experienced on December 7.

Finally, Coffman rejects Allen's attempt to show lack of price impact by using Honeywell, 3M, and Dow's lack of price reaction during the class period as compared to their increase in implied volatility and price reaction at the end of and after the class period. Coffman Rep. at 46 ¶ 92. First, Coffman notes that Honeywell, 3M, and Dow's lack of price impact during the class period is unsurprising, when Allen herself admits there was no correlative news during the class period. *Id.* at 47 ¶ 93. He also argues that Dow and Honeywell's price reactions after the class period show that stock prices can and do react negatively to asbestos-related developments. *Id.* at 48 ¶ 97. Coffman further argues that a number of companies went bankrupt during the class period because of asbestos liability, a fact Allen acknowledges. *Id.* at 49 ¶ 98; Pl. Ex. 15, Dkt. No. 591, Pl. App. 114. Coffman explains that Allen is relying on a very small sample size of companies to support her sweeping conclusion. *Id.* at 49–50 ¶ 99. He argues that Allen's own report strongly supports the view that the market viewed Halliburton as an asbestos company starting on December 7, and not before, which shows that the disclosure of the Baltimore verdict on that date caused the market price to reflect that investors were no longer relying on Halliburton's misrepresentations.<sup>45</sup> Coffman Rep. at 53 ¶ 108.

The Court finds that Halliburton has not met its burden of showing lack of price impact with respect to the announcement of the Baltimore verdict on December 7th. Although the Court finds that at least some of Halliburton's stock price decline on that date is likely attributable to uncertainty in the asbestos environment that also impacted other companies with asbestos exposure, Halliburton has not demonstrated that uncertainty caused the entirety of Halliburton's

---

<sup>45</sup> Coffman notes that Allen's Chow test shows that after a series of negative litigation outcomes, Halliburton's stock became correlated with the asbestos index, *i.e.*, the market started treating Halliburton as an asbestos company. Coffman Rep. at 53 ¶ 108. A Chow test looks at whether relationships between variables have changed, and here, the test indicated that the market started treating Halliburton as a company with material asbestos liabilities. Hr'g Tr. at 186:24-187:9.


substantial price decline. *See* Hr’g Tr. at 183:15-21 (Coffman explaining that the change in the asbestos climate generally was only a small factor in Halliburton’s price decline). The Court finds that the price impact on December 7 likely reflected the market’s view of Halliburton’s prior representations regarding its asbestos liability *and* increased uncertainty in the asbestos environment. *See* Hr’g Tr. at 190:18-191:3 (Coffman explaining that economic common sense suggests that a corrective disclosure can both impact Halliburton’s price, along with the share prices of other companies). Allen’s conclusion with respect to December 7 is that there was no price impact *as to the alleged misrepresentation*. However, at the evidentiary hearing, Allen struggled to articulate why the price impact on December 7 could not be caused by the market’s realization that Halliburton knew it faced increased asbestos exposure and had concealed that fact before December 7, as alleged by the Fund. Hr’g Tr. at 126:3-127:6. The substantial price decline on December 7 and Coffman’s event study show price impact as to the corrective disclosure on that date, and Halliburton has not shown that the other factors it relies upon show otherwise.

### III. CONCLUSION

The Court **GRANTS in part** Plaintiffs’ Motion for Class Certification, only with respect to the alleged corrective disclosure of December 7, 2001. The Motion for Class Certification is **DENIED** as to the other five corrective disclosures on which Plaintiffs rely.

**SO ORDERED.**

July 25, 2015.

  
BARBARA M. G. LYNN  
UNITED STATES DISTRICT JUDGE  
NORTHERN DISTRICT OF TEXAS



# Exhibit 67



Slip Copy  
Slip Copy, 2006 WL 845161 (M.D.Fla.)  
(Cite as: 2006 WL 845161 (M.D.Fla.))

Page 1

### Motions, Pleadings and Filings

Only the Westlaw citation is currently available.

United States District Court,  
M.D. Florida.  
In re TECO ENERGY, INC. SECURITIES  
LITIGATION,  
**No. 8:04-CV-1948-T-27EAJ.**

March 30, 2006.

[Darren J. Robbins](#), Lerach Coughlin Stoia & Robbins Llp, [William S. Lerach](#), Lerach Coughlin Stoia Geller Rudman & Robbins Llp, San Diego, CA, [David J. George](#), Lerach Loughlin, Jack Reise, [Stephen Richard Astley](#), Lerach Coughlin Stoia Geller Rudman & Robbins Llp, [Kenneth J. Vianale](#), Vianale & Vianale Llp, Boca Raton, FL, [David A. Rosenfeld](#), Lerach Coughlin Stoia Geller Rudman & Robbins Llp, [Samuel H. Rudman](#), Geller Rudman, Plc, Melville, NY, for Plaintiffs.

Jade L. Kurtz, pro se.

Robert A. Kurtz, pro se.

Frank R. Crisanto, pro se.

David R. Withers, pro se.

[Diane Knox](#), [Richard A. Rosen](#), Paul, Weiss, Rifkind, Wharton & Garrison Llp, [Steven B. Rosenfeld](#), New York, NY, [Tracy A. Nichols](#), Holland & Knight Llp, Miami, FL, for Defendants.

### *ORDER*

[WHITEMORE](#), J.

\*1 BEFORE THE COURT are Defendants' Motion to Dismiss (Dkt.64), Plaintiffs' Opposition (Dkt.76), Defendants' Supplemental Memorandum of Law (Dkt.80), and Plaintiffs' Response (Dkt.83). Upon consideration, Defendants' Motion to Dismiss (Dkt.64) is GRANTED in part and DENIED in part.

Plaintiffs filed their Consolidated Class Action Complaint alleging violations of Section 10(b) of the Securities Exchange Act of 1934 ("Exchange Act")

and Rule 10b-5 promulgated thereunder against all Defendants and Section 20(a) of the Exchange Act against the individual Defendants. [\[FN1\]](#) (Dkt.59, Compl., ¶ 1). Defendants, TECO Energy, Inc. ("TECO"), Robert D. Fagan and Gordon L. Gillette, move to dismiss the Complaint pursuant to [Fed.R.Civ.P. 12\(b\)\(6\)](#) based on Plaintiffs' failure to plead loss causation and failure to state a fraud claim under Section 10(b). (Dkt.64).

[FN1](#). Section 10(b) of the Exchange Act prohibits the use or employment of any manipulative or deceptive device in connection with the purchase or sale of any security in contravention of Securities Exchange Commission rules and regulations. 15 U.S.C. § 78j(b). Rule 10b-5 prohibits the making of any "untrue statement of a material fact" or the omission of a material fact "necessary in order to make the statements made, in light of the circumstances under which they were made, not misleading." [17 C.F.R. § 240.10b-5](#). Pursuant to Section 20(a) of the Exchange Act, liability may be imposed on a "controlling person" where a securities violation is found. [15 U.S.C. § 78t\(a\); Brown v. Enstar Group, Inc.](#), 84 F.3d 393, 395-97 (11th Cir.1996).

### *Factual Background*

Plaintiffs are NECA-IBEW Pension Fund, Monroe County Employees Retirement System, John Marder, and Charles Korpak, individually and on behalf of a proposed class of persons who purchased publicly traded securities of TECO between October 30, 2001 and February 4, 2003 ("the class period"). (Dkt.59, ¶ 1, 16). TECO is a public utility holding company for regulated utilities and other unregulated businesses. (Dkt.59, ¶ 17). TECO owns no operating assets but holds all of the common stock of its regulated operating subsidiary, Tampa Electric Company, and other non-regulated subsidiaries. (Dkt.59, ¶ 17). Defendant, Robert D. Fagan, is former Chief Executive Officer and President of TECO and Chairman of TECO's Board of Directors. (Dkt.59, ¶ 18). Defendant, Gordon L. Gillette, is TECO's former Chief Financial Officer. (Dkt.59, ¶ 19).

In essence, Plaintiffs allege that Defendants engaged

Slip Copy  
Slip Copy, 2006 WL 845161 (M.D.Fla.)  
(Cite as: 2006 WL 845161 (M.D.Fla.))

Page 2

in a fraudulent scheme to misrepresent TECO's financial condition and to artificially inflate TECO's stock price by misrepresenting its financial results, success, and prospects, including misrepresentations and omissions regarding TECO's: (1) abandonment of its prior business model; (2) liability for its multi-billion dollar power plant projects; (3) inability to sell or transmit power from its merchant energy power plants; (4) exposure to Enron's demise; and (5) impossibility of maintaining its dividend. (Dkt.76, pp. 1, 24). In addition, Plaintiffs allege that Defendants falsified TECO's financial results in 2001 and 2002 through improper accounting practices. (Dkt.76, p. 24). In large part, Plaintiffs rely on analysts' reports that they allege revealed Defendants' fraud, thereby removing the fraud-based inflation in TECO's stock price. (Dkt.76, pp. 12-13, 24-26).

#### *Applicable Standards*

A court may grant a motion to dismiss "only when the defendant demonstrates beyond doubt that the plaintiff can prove no set of facts in support of his claim which would entitle him to relief." Chepstow Ltd. v. Hunt, 381 F.3d 1077, 1080 (11th Cir.2004) (internal quotation and citation omitted). The court will accept as true the factual allegations in the complaint and will view them in a light most favorable to the nonmoving party. *Id.* In considering a motion to dismiss, the court is generally confined to examining the four corners of the complaint, but the court may take judicial notice of relevant documents publicly filed with the Securities Exchange Commission ("SEC"). Bryant v. Avado Brands, Inc., 187 F.3d 1271, 1287 (11th Cir.1999).

\*2 To state a cause of action under Section 10(b) or Rule 10b-5, plaintiffs are required to allege: "(1) a misstatement or omission (2) of a material fact (3) made with scienter (4) upon which the plaintiff relied (5) that proximately caused the plaintiff's loss." Theoharous v. Fong, 256 F.3d 1219, 1224 (11th Cir.2001) (citation omitted). In passing the Private Securities Litigation Reform Act of 1995 ("PSLRA"), Congress imposed a heightened pleading requirement for claims alleging violations of Section 10(b) and Rule 10b-5. 15 U.S.C. § 78u-4(b)(1)(2); Harris v. Ivax Corp., 182 F.3d 799, 803 (11th Cir.1999). However, the elements of proximate causation and economic loss are not subject to the heightened pleading requirement and must only be plead in accordance with Fed.R.Civ.P. 8(a)(2), which requires a "short and plain statement of the claim showing that the pleader is entitled to relief." Dura Pharms., Inc. v. Broudo, 544 U.S. 336, 125 S.Ct. 1627, 1634, 161 L.Ed.2d 577 (2005) (quoting Fed.R.Civ.P. 8(a)(2)).

Notwithstanding, "the short and plain statement must provide the defendant with fair notice of what the plaintiff's claim is and the grounds upon which it rests." Dura Pharms., Inc., 125 S.Ct. at 1634 (quoting Conley v. Gibson, 355 U.S. 41, 47, 78 S.Ct. 99, 2 L.Ed.2d 80 (1957)) (internal quotations omitted).

The threshold is "exceedingly low" for a complaint to survive a motion to dismiss for failure to state a claim. United States v. Baxter Int'l, Inc., 345 F.3d 866, 881 (11th Cir.2003). The PSLRA did not alter the required presumption that the court "give Plaintiffs, not Defendants, the benefit of every favorable inference that can be drawn from their allegations." In re Sykes Enters., Inc., No. 8:00-CV-212-RAL, 2001 WL 964160, at \*2 (M.D.Fla. Mar.7, 2001) (citations omitted).

#### *Discussion*

##### *I. Loss Causation*

Pursuant to the PSLRA, a plaintiff has the burden of proving that each alleged misrepresentation or omission proximately caused the plaintiff's economic loss. [FN2] 15 U.S.C. § 78u-4(b)(4); Dura Pharms., Inc., 125 S.Ct. at 1633. "[T]o establish loss causation, a plaintiff must allege that the subject of the fraudulent statement or omission was the cause of the actual loss suffered, i.e., that the misstatement or omission concealed something from the market that, when disclosed, negatively affected the value of the security." Lentell v. Merrill Lynch & Co., 396 F.3d 161, 173 (2d Cir.2005) (internal quotations and citation omitted). Even under the liberal notice pleading standard, "it should not prove burdensome for a plaintiff who has suffered an economic loss to provide a defendant with some indication of the loss and the causal connection that the plaintiff has in mind." Dura Pharms., Inc., 125 S.Ct. at 1634.

FN2. See also Robbins v. Koger Props., Inc., 116 F.3d 1441, 1447 (11th Cir.1997). In Robbins, the Eleventh Circuit provided the following standard for proving loss causation:

To prove loss causation, a plaintiff must show that the untruth was in some reasonably direct, or proximate, way responsible for his loss. If the investment decision is induced by misstatements or omissions that are material and that were relied on by the claimant, but are not the proximate reason for his pecuniary loss, recovery under the Rule is not permitted. In

Slip Copy  
 Slip Copy, 2006 WL 845161 (M.D.Fla.)  
 (Cite as: 2006 WL 845161 (M.D.Fla.))

Page 3

other words, loss causation describes the link between the defendant's misconduct and the plaintiff's economic loss. Because market responses, such as stock downturns, are often the result of many different, complex, and often unknowable factors, the plaintiff need not show that the defendant's act was the sole and exclusive cause of the injury he has suffered; he need only show that it was "substantial," i.e., a significant contributing cause.

Robbins, 116 F.3d at 1447. This standard is not changed by *Dura*.

In *Dura*, the U.S. Supreme Court held that loss causation may not be established by simply alleging a stock was purchased at an artificially inflated price. *Dura Pharms., Inc.*, 125 S.Ct. at 1631-32, 1364. Rather, to sufficiently plead loss causation, a plaintiff must allege a disclosure or revelation of truth about a defendant's prior misstatement or omission that is in some way connected with a stock price drop. [FN3] *Id.* at 1634 (motion to dismiss granted based on plaintiffs' failure to plead loss causation where plaintiffs failed to allege that stock prices fell after "the truth became known"). Establishing a connection between a drop in stock price and the disclosure of the "truth" about a defendant's previous misstatement or omission is essential in pleading loss causation, even if that connection may be made in a short and plain statement. *Id.* at 1634. If a drop in stock price occurs before a defendant's fraud is revealed or the truth becomes known, the damages associated with the drop in stock price necessarily cannot be connected to the alleged fraud. *In re Daou Sys., Inc.*, 411 F.3d at 1026-27. Similarly, if a plaintiff sells his stock before a corrective disclosure is made, the plaintiff cannot show that his losses resulted from previously undisclosed fraud. *Davidco Investors, LLC v. Anchor Glass Container Corp.*, No. 8:04-CV-2561-T-24EAJ, 2006 WL 547989, at \*10 (M.D.Fla. Mar.6, 2006) (citing *Dura Pharms., Inc.*, 125 S.Ct. at 1631). Moreover, generalized, vague or overbroad allegations regarding the existence of a disclosure or revelation of fraud that is merely alleged to have been connected to a drop in stock price will not suffice to put a defendant on notice of loss causation claims. *Dura Pharms., Inc.*, 125 S.Ct. at 1634.

[FN3. See also *In re Daou Sys., Inc.*, 411 F.3d 1006, 1025-27 (9th Cir.2005) (motion to dismiss denied where plaintiffs sufficiently plead loss causation based on disclosures regarding defendants' "true

financial health" and a resulting drop in stock price); *In re Sawtek, Inc.*, No. 6:03-CV-294-ORL-31-DAB, 2005 WL 2465041, at \*11-12 (M.D.Fla. Oct.6, 2005) (motion to dismiss granted based on plaintiffs' failure to plead a connection between defendants' alleged fraud and the losses allegedly sustained as a result).

\*3 Defendants argue that Plaintiffs have failed to sufficiently plead loss causation as they have failed to show a causal connection between the alleged fraud and their economic losses. (Dkt.64, p. 1). Specifically, Defendants argue that Plaintiffs have failed to show a corrective disclosure or a materialization of a previously concealed risk that caused the drop in Defendants stock prices that ultimately led to their losses. (Dkt.64, p. 1). Moreover, Defendants argue that Plaintiffs' Complaint alleges that "the truth" first began to "leak out" on September 3, 2002, when information appeared in the market via analysts' reports to correct alleged misstatements or omissions by Defendants. (Dkt.64, pp. 2, 13-14). Defendants argue that, even assuming that the truth about Defendants' financial condition was revealed, the fact that the information first entered the market on September 3, 2002, more than halfway through the October 31, 2001 to February 4, 2003 class period, precludes any claim for market losses prior to September 3, 2002, as those losses cannot be causally connected to the September 3, 2002 revelation of alleged fraud. [FN4] (Dkt.64, pp. 2, 13-17). Defendants also argue that Plaintiffs' Complaint fails to sufficiently plead loss causation as to market losses incurred after September 3, 2002 as Plaintiffs do not connect any information disclosed in the market to any misstatement or omission by Defendants. (Dkt.64, pp. 2, 18-29). Further, Defendants argue that Plaintiffs' Complaint fails to sufficiently plead loss causation as to its alleged improper accounting practices because there are no causally connected class-period losses associated with the relevant practices alleged as improper in the Complaint. (Dkt.64, pp. 2, 29-31).

[FN4. Although Defendants' argument that the class period should begin no earlier than September 3, 2002 arguably has merit, the Court will not address class issues in ruling on the instant motion.

Despite the parties' lengthy arguments, the only relevant issue for consideration on Defendants' motion to dismiss on loss causation grounds is whether, when and how Defendants' alleged

Slip Copy  
 Slip Copy, 2006 WL 845161 (M.D.Fla.)  
 (Cite as: 2006 WL 845161 (M.D.Fla.))

Page 4

misstatements and omissions were revealed to the market, thereby causing a drop in stock prices. [\[FNS\]](#) In arguing that they have sufficiently pleaded loss causation, Plaintiffs point to ¶¶ 167-266 and 288 of their Complaint which allege misstatements, omissions, and improper accounting practices, and ¶¶ 12, 206-240, 290-291, 296-298 which allege loss causation. (Dkt.76, pp. 24-26). Plaintiffs essentially suggest to the Court that sufficient allegations of loss causation can be gleaned from those 150 paragraphs of the Complaint. Even applying a notice pleading standard, however, those broad allegations will not suffice. Notwithstanding the 164 page, 316 paragraph Complaint, and drawing all favorable inferences in Plaintiffs' favor, the Court cannot find sufficient allegations of loss causation required by *Dura*.

[FN5.](#) In their opposition and response, Plaintiffs spend an inordinate amount of time arguing their position on the other elements of their fraud claim. These arguments are neither relevant nor helpful as Defendants' seek dismissal based on failure to plead loss causation and do not challenge the other elements. Further, the parties spend an inordinate of time arguing about the requirements for proving loss causation. Defendants assert that Plaintiffs must allege a "corrective disclosure" or "a materialization of a concealed risk." Plaintiffs assert that they must only show "some indication" of a causal connection via a "revelation" of "the relevant truth" or Defendants' "true financial condition." Regardless of the words or labels used, the requirements are the same. Under *Dura*, it is not enough for Plaintiffs to allege an inflated stock price. Rather, Plaintiffs must allege that some truth was disclosed in the market that revealed prior misstatements or omissions--fraud--by Defendants that is causally connected to their losses. [Dura Pharms., Inc., 125 S.Ct. at 1634.](#)

Plaintiffs argue that "the truth about TECO's financial condition gradually began to emerge in September 2002 when analysts began to openly question TECO's future earnings and dividend payment prospects." (Dkt.76, p. 25). Plaintiffs further argue that the Complaint "links the emergence of the truth of TECO's true financial condition and business operations with the removal of the fraud-based inflation in the stock price, thereby showing the causal connection between the fraudulent scheme and plaintiffs' economic loss." (Dkt.76, p. 25). In support,

Plaintiffs point to eight purported revelations of the truth on six different dates:

\*4 (1) Plaintiffs allege that TECO's stock price dropped on September 3, 2002 after a severe ratings cut by three analysts who issued reports questioning TECO's future earnings and dividend payment prospects. (Dkt.59, ¶ 206). Plaintiffs allege that through these analysts' reports, "news leaked out that TECO was not doing nearly as well as prior representations." (Dkt.59, ¶ 296).

(2) Plaintiffs allege that in response to a September 23, 2002 press release by TECO, an analyst reported that TECO was "overstating its earnings power," "understating its true capital needs," and needed "to raise an additional \$700 mil." in financing, which led the analyst to predict that TECO would not be able to fund its dividend, making it a "misleading indicator of value and a poor proxy for the underlying cashflow economics of the business." (Dkt.59, ¶ 210).

(3) Plaintiffs allege that on October 8, 2002, an analyst questioned TECO's disclosures and plan to issue 15 million shares of common stock. (Dkt.59, ¶ 221). The analyst noted that, "Since it released its updated financial plan on 9/23, management has consistently stated that it requires no additional equity. This announcement could indicate that the company's problems are worse than initially indicated by [TECO]." (Dkt.59, ¶ 221).

(4) Plaintiffs allege that on October 11, 2002, TECO filed a prospectus indicating restructuring associated with various projects. (Dkt.59, ¶¶ 222-223).

(5) Plaintiffs allege that Merrill Lynch issued a report on January 24, 2003, days after TECO reported its 2002 earnings, indicating that TECO had reached "the end of the good times," lowering its 2003 earnings per share estimates for TECO, and predicting "full-year losses" in 2003. (Dkt.59, ¶ 237). Plaintiffs also allege a similar report was issued in mid-January 2003 by SalomonSmithBarney. (Dkt.59, ¶ 238).

(6) Plaintiffs allege that on February 4, 2003, Moody's Investors Service downgraded Panda Funding Corp.'s \$99 million of senior secured debt, and these dividends were the only source of cash to pay back debt to TECO. (Dkt.59, ¶ 239).

Finally, Plaintiffs allege that as of February 4, 2003, "TECO's house of cards finally collapsed as even more of the truth emerged." (Dkt.59, ¶ 239). Plaintiffs allege that based on these revelations of "truth" and with TECO's "true financial condition exposed," the stock price dropped to \$12.78 as of February 5, 2003. (Dkt. 76, p. 26; Dkt. 59, ¶ 240).



Slip Copy  
 Slip Copy, 2006 WL 845161 (M.D.Fla.)  
 (Cite as: 2006 WL 845161 (M.D.Fla.))

Page 5

Although the "revelations" referenced by Plaintiffs suggest that analysts were pessimistic regarding TECO's future, the information contained in the purported revelations does not identify, reveal or correct any prior misstatement, omission, or improper accounting practice by Defendants. In fact, none of the purported revelations indicate or establish that the changes occurring with TECO were remotely associated with prior fraudulent conduct. Specifically, the analyst reports on September 3, 2002, September 23, 2002, October 8, 2002, and January 23, 2003 address ratings cuts, opinions, and predictions regarding TECO's stock value but do not reference any misstatements, omissions, or accounting practices by Defendants as the reason for the bleak forecasts or changes in market conditions. Similarly, Defendants' October 11, 2002 prospectus and the ratings cuts and downgrade of debt cited by Plaintiffs provide no connection to prior misstatements, omissions, or improper accounting practices. [\[FN6\]](#) The opinions, predictions, and generalized statements offered by Plaintiffs as "revelations" of the "truth" regarding TECO's financial status, without more, are not sufficient to establish loss causation.

[FN6.](#) Plaintiffs do not allege that Defendants filed any restatements during the relevant period.

\*5 Moreover, the revelations relied on by Plaintiffs do not sufficiently establish a connection between the specific fraudulent activity alleged and a drop in stock prices. Even assuming that the revelations gave some indication of prior misstatements, omissions, or improper accounting practices, Plaintiffs have not sufficiently alleged that those revelations were related to the fraudulent scheme alleged in the Complaint. To prove loss causation, Plaintiffs must allege that a misstatement or omission by Defendants concealed something from the market and that when information related to that fraud was disclosed, the value of their securities were effected. See [Lentell, 396 F.3d at 173](#). In addition to failing to reference any prior misstatement, omission, or improper accounting practice, the information contained in the eight revelations relied on by Plaintiffs does not specifically relate to the issues involved in the alleged fraudulent scheme. See [Barr v. Matria Healthcare, Inc., 324 F.Supp.2d 1369, 1380 \(N.D.Ga.2004\)](#) (revelation or disclosure of prior misrepresentation must discuss the subject matter of the alleged fraud to satisfy loss causation).

Even applying a notice pleading standard, drawing all inferences in Plaintiffs' favor, and assuming the information described above was revealed, Plaintiffs have not sufficiently plead a connection between Defendants' alleged fraud and their economic loss. See [Dura Pharms., Inc., 125 S.Ct. at 1634; Robbins, 116 F.3d at 1447](#). Plaintiffs' broad, vague allegations regarding the existence of a disclosure or revelation of fraud or improper accounting practices are not sufficient to put Defendants on notice of the Plaintiffs' loss causation claims. See [Dura Pharms., Inc., 125 S.Ct. at 1634](#). In sum, Plaintiffs have not sufficiently alleged that Defendants' fraud, as opposed to poor market conditions, was the proximate cause of TECO's stock price decline. Accordingly, Defendants' motion to dismiss Plaintiffs' Complaint for failure to plead loss causation is granted.

## *II. Fraud vs. Mismanagement*

As an additional reason for dismissal, Defendants argue that Plaintiffs' Complaint fails to state a claim under Section 10(b) because the alleged conduct amounts to mismanagement, not fraud. (Dkt.64, pp. 31-34). While Defendants are correct that Section 10(b) claims are only actionable if they involve deception or manipulation, the Court is not prepared to dismiss Plaintiffs' Complaint on these grounds. See [Santa Fe Indus., Inc. v. Green, 430 U.S. 462, 473-74, 97 S.Ct. 1292, 51 L.Ed.2d 480 \(1977\)](#). Plaintiffs, despite their lengthy filings, have failed to provide adequate argument regarding this issue. Defendants, on the other hand, have chosen not to challenge the elements of Plaintiffs' fraud claim (other than loss causation), yet seek dismissal for failure to state a claim.

Notwithstanding, upon review of Plaintiffs' Complaint, accepting all of Plaintiffs' allegations as true, the Court finds that Plaintiffs have set forth allegations of deception and manipulation sufficient to support a fraud claim under Section 10(b), apart from the issues related to loss causation. Therefore, Defendants' motion to dismiss Plaintiffs' Complaint for failure to state a claim under Section 10(b) is denied.

\*6 Accordingly, it is ORDERED AND ADJUDGED that Defendants' Motion to Dismiss (Dkt.64) is GRANTED IN PART and DENIED IN PART. Defendants' Motion to Dismiss Plaintiffs' Complaint for failure to plead loss causation is GRANTED. Defendants' Motion to Dismiss Plaintiffs' Complaint for failure to state a claim under Section 10(b) is

Slip Copy  
Slip Copy, 2006 WL 845161 (M.D.Fla.)  
(Cite as: 2006 WL 845161 (M.D.Fla.))

Page 6

DENIED. Plaintiffs' Consolidated Class Action Complaint (Dkt.59) is DISMISSED WITHOUT PREJUDICE.

DONE AND ORDERED.

Slip Copy, 2006 WL 845161 (M.D.Fla.)

**Motions, Pleadings and Filings** ([Back to top](#))

- [2005 WL 2318159](#) (Trial Motion, Memorandum and Affidavit) Defendants' Motion to Dismiss and Incorporated Memorandum of Law (Jul. 25, 2005)
- [2005 WL 2318161](#) (Trial Motion, Memorandum and Affidavit) Defendants' Motion to Dismiss and Incorporated Memorandum of Law (Jul. 25, 2005)
- [2005 WL 2318163](#) (Trial Motion, Memorandum and Affidavit) Defendants' Motion to Dismiss and Incorporated Memorandum of Law (Jul. 25, 2005)
- [8:04cv01948](#) (Docket) (Aug. 25, 2004)
- [2004 WL 2179889](#) (Trial Pleading) Complaint for Violation of the Federal Securities Laws (Aug. 24, 2004)

END OF DOCUMENT



# Exhibit 68

330 F.R.D. 439

United States District Court, W.D. Texas, Austin Division.

John ROONEY, Individually and on Behalf of All Others Similarly Situated, Plaintiff,

v.

EZCORP, INC. and Mark E. Kuchenrither, Defendants.

Cause No.: A-15-CA-00608-SS

|

Signed 02/19/2019

### Synopsis

**Background:** Investor brought putative securities fraud class action under, inter alia, § 10(b) and Rule 10b-5, against corporation and its officer, alleging that defendants made material misrepresentations to artificially inflate price of corporation's stock, which resulted in financial losses following release of corporation's restated financial reports. Investor moved to certify class.

**Holdings:** The District Court, [Sam Sparks](#), Senior District Judge, held that:

- [\[1\]](#) numerosity requirement was satisfied;
- [\[2\]](#) commonality requirement was satisfied;
- [\[3\]](#) typicality requirement was satisfied;
- [\[4\]](#) adequacy requirement was satisfied;
- [\[5\]](#) predominance requirement was satisfied; and
- [\[6\]](#) superiority requirement was satisfied.

Motion granted.

**Procedural Posture(s):** Motion to Certify Class.

## West Headnotes (23)

**[1] Federal Civil Procedure** 🔑 Evidence; pleadings and supplementary material

Plaintiffs seeking to certify a class under the rule governing class certification bear the burden of establishing the prerequisites to certification have been met. [Fed. R. Civ. P. 23](#).

**[2] Federal Civil Procedure** 🔑 Stockholders, investors, and depositors

Proposed class was so numerous so as to render joinder impracticable, as required for certification of proposed class consisting of those who purchased or otherwise acquired corporation's class A common stock within certain time period and who were damaged as a result, in investor's putative class action against corporation and its officer alleging, inter alia, that defendants made material misrepresentations to artificially inflate corporation's stock price in violation of § 10(b) and Rule 10b-5, even though exact number of class members was unknown; corporation had more than 50 million shares of Class A common stock outstanding during class period, and the average weekly trading volume on stock exchange during class period was roughly 2.7 million shares. Securities Exchange Act of 1934, § 10(b), [15 U.S.C.A. § 78j](#); [17 C.F.R. § 240.10b-5](#); [Fed. R. Civ. P. 23\(a\)\(1\)](#).

**[3] Federal Civil Procedure** 🔑 Stockholders, investors, and depositors

Commonality requirement was satisfied, as required for certification of proposed class consisting of those who purchased or otherwise acquired corporation's class A common stock within certain time period and who were damaged as a result, in investor's putative class action against corporation and its officer alleging, inter alia, that defendants made material misrepresentations to artificially inflate price of corporation's stock in violation of § 10(b) and Rule 10b-5; investor alleged uniform representations and omissions made to class concerning accuracy of corporations' publicly reported financial results, and allegations implicated multiple questions common to class, including whether defendants' statements or omissions violated federal securities law. Securities Exchange Act of 1934, § 10(b), [15 U.S.C.A. § 78j](#); [17 C.F.R. § 240.10b-5](#); [Fed. R. Civ. P. 23\(a\)\(2\)](#).

**[4] Federal Civil Procedure** 🔑 Stockholders, investors, and depositors

Investor, named plaintiff, established that his claims were typical of those of proposed class, as required for certification of class consisting those who purchased or otherwise acquired corporation's class A common stock within certain time period and who

were damaged as a result, in putative class action against corporation and its officer alleging, inter alia, that defendants made material misrepresentations to artificially inflate corporation's stock price in violation of § 10(b) and Rule 10b-5; all putative class members and investor allegedly purchased corporation's stock during class period at prices inflated by defendants' misstatements and omissions, investor presented classwide legal theory, and investor did not appear to be subject to any unique defense inapplicable to class as a whole. Securities Exchange Act of 1934, § 10(b), 15 U.S.C.A. § 78j; 17 C.F.R. § 240.10b-5; Fed. R. Civ. P. 23(a)(3).

**[5] Federal Civil Procedure** 🔑 Representation of class; typicality; standing in general

The purpose of the adequacy requirement for class certification is to uncover conflicts of interest between named parties and the class they seek to represent. Fed. R. Civ. P. 23(a)(4).

**[6] Federal Civil Procedure** 🔑 Representation of class; typicality; standing in general

In order to meet the adequacy requirement for class certification, the plaintiff must show that he is willing and able to vigorously prosecute the interests of the class through qualified counsel. Fed. R. Civ. P. 23(a)(4).

**[7] Federal Civil Procedure** 🔑 Representation of class; typicality; standing in general

Although class representatives need not be legal scholars and are entitled to rely on counsel, in order to satisfy the adequacy requirement for class certification, the plaintiff must at least know more than that they were involved in a bad business deal. Fed. R. Civ. P. 23(a)(4).

**[8] Federal Civil Procedure** 🔑 Stockholders, investors, and depositors

The Private Securities Litigation Reform Act (PSLRA) provisions governing the selection of class representatives do not affect the inquiry into whether a proposed class representative has met adequacy requirement of the rule governing class certification. Securities Exchange Act of 1934, § 21D, as amended, 15 U.S.C.A. § 78u-4; Fed. R. Civ. P. 23(a)(4).

1 Case that cites this headnote

**[9] Federal Civil Procedure** 🔑 Stockholders, investors, and depositors

Residency of investor, the named plaintiff, in Ireland, did not render investor an inadequate class representative, for purposes of certification of proposed class consisting of those who purchased or otherwise acquired corporation's class A common stock within certain time period and who were damaged as a result, in putative class action against corporation and its officer alleging, inter alia, that defendants made material misrepresentations to artificially inflate price of corporation's stock in violation of § 10(b) and Rule 10b-5; investor was able to easily correspond with counsel through email, telephone, and video chat, and there was little reason to believe investor's residence posed obstacle to his ability to appear in court as needed since investor declared he would be available to travel. Securities Exchange Act of 1934, § 10(b), 15 U.S.C.A. § 78j; 17 C.F.R. § 240.10b-5; Fed. R. Civ. P. 23(a)(4).

**[10] Federal Civil Procedure** 🔑 Stockholders, investors, and depositors

Investor, named plaintiff, demonstrated sufficient knowledge regarding subject matter of litigation to serve as adequate class representative, as required for certification of proposed class consisting of those who purchased or otherwise acquired corporation's class A common stock within certain time period and who were damaged as a result, in putative class action against corporation and its officer alleging, inter alia, that defendants made material misrepresentations to artificially inflate corporation's stock price in violation of § 10(b) and Rule 10b-5; investor demonstrated that he knew more than that he was involved in a bad business deal, including by testifying at deposition that he believed corporation violated federal securities law by misleadingly classifying some of company's loans in order to report profit. Securities Exchange Act of 1934, § 10(b), 15 U.S.C.A. § 78j; 17 C.F.R. § 240.10b-5; Fed. R. Civ. P. 23(a)(4).

**[11] Federal Civil Procedure** 🔑 Stockholders, investors, and depositors

Investor, named plaintiff, met the adequacy requirement, as required for certification of proposed class consisting of those who purchased or otherwise acquired corporation's class A common stock within certain time period and who were damaged as a result, in putative class action against corporation and its officer alleging that defendants made material misrepresentations to artificially inflate price of corporation's stock in violation of § 10(b) and Rule 10b-5; district court was not aware any conflicts between investor and members of proposed class, as best court could tell investor's interests were aligned with those of the class as a whole, and investor demonstrated he was both willing and able to vigorously prosecute interests of class through qualified counsel. Securities Exchange Act of 1934, § 10(b), 15 U.S.C.A. § 78j; 17 C.F.R. § 240.10b-5; Fed. R. Civ. P. 23(a)(4).

**[12] Federal Civil Procedure** 🔑 Common interest in subject matter, questions and relief; damages issues

The inquiry to determine whether the predominance requirement for class certification is satisfied tests whether the proposed class is sufficiently cohesive to warrant adjudication by representation. [Fed. R. Civ. P. 23\(b\)\(3\)](#).

**[13] Federal Civil Procedure** 🔑 Stockholders, investors, and depositors

In securities law cases, predominance, for purposes of class certification, often hinges upon whether or not class members will need to individually demonstrate they relied upon a misrepresentation or omission made by the defendant. [Fed. R. Civ. P. 23\(b\)\(3\)](#).

1 Case that cites this headnote

**[14] Federal Civil Procedure** 🔑 Stockholders, investors, and depositors

If the class members in a securities fraud class action cannot establish they are entitled to rely on the presumption that an investor who buys or sells stock at the market price has relied on any public material misrepresentations, or if the defendant successfully rebuts it, class members must individually establish reliance, and in practice, such individual inquiries almost inevitably prevent the plaintiff from establishing predominance requirement for class certification by overwhelming whatever common questions may exist. [Fed. R. Civ. P. 23\(b\)\(3\)](#).

1 Case that cites this headnote

**[15] Federal Civil Procedure** 🔑 Stockholders, investors, and depositors

**Securities Regulation** 🔑 Fraud on the market

To establish that the class in a securities class action is entitled to rely on the presumption that an investor who buys or sells stock at the market price has relied on any public material misrepresentations, for purposes of the predominance requirement for class certification, the plaintiff must demonstrate: (1) the alleged misrepresentations were publicly known; (2) the alleged misrepresentations were material; (3) the stock traded in an efficient market; and (4) putative class members traded the stock between the time the misrepresentations were made and when the truth was revealed. [Fed. R. Civ. P. 23\(b\)\(3\)](#).

1 Case that cites this headnote

**[16] Federal Civil Procedure** ➡ Stockholders, investors, and depositors**Securities Regulation** ➡ Fraud on the market

Investor, named plaintiff, established that proposed class was entitled to rely on presumption that an investor who buys or sells stock at market price has relied on any public material misrepresentations, for purposes of predominance requirement for certification of class consisting of those who purchased and acquired corporation's class A common stock and who were damaged as result, in putative class action against corporation and officer alleging, inter alia, material misrepresentations to artificially inflate corporation's stock price in violation of § 10(b) and Rule 10b-5; alleged misrepresentations were publicly known, alleged misrepresentations were presumed to be material at class certification stage, corporation's stock traded in efficient market, and proposed class traded the stock between time misrepresentations were made and when they were corrected. Securities Exchange Act of 1934, § 10(b), 15 U.S.C.A. § 78j; 17 C.F.R. § 240.10b-5; Fed. R. Civ. P. 23(b)(3).

1 Case that cites this headnote

**[17] Securities Regulation** ➡ Fraud on the market

The presumption, that an investor who buys or sells stock at the market price has relied on any public material misrepresentations, relies upon indirect evidence of stock price impact to establish reliance on misrepresentations; direct evidence that the misrepresentation did not affect the price of the stock severs that link.

5 Cases that cite this headnote

**[18] Federal Civil Procedure** ➡ Stockholders, investors, and depositors**Securities Regulation** ➡ Presumptions and burden of proof

Defendants, corporation and officer, did not rebut presumption, as would preclude investor from relying on presumption that an investor who buys or sells stock at market price has relied on any public material misrepresentations, for purposes of establishing predominance requirement for certification of proposed class consisting those who purchased or acquired corporation's class A common stock during certain time period and who were damaged as result, in putative class action alleging, inter alia, that defendants made material misrepresentations to artificially inflate corporation's stock price in violation of § 10(b) and Rule 10b-5, despite contention that there was no statistically significant price adjustment following certain corrective disclosure dates; defendants did not point to evidence that alleged misrepresentations did not affect stock price. Securities



Exchange Act of 1934, § 10(b), 15 U.S.C.A. § 78j; 17 C.F.R. § 240.10b-5; Fed. R. Civ. P. 23(b)(3).

5 Cases that cite this headnote

**[19] Federal Civil Procedure** 🔑 Stockholders, investors, and depositors  
**Securities Regulation** 🔑 Presumptions and burden of proof

Defendants can rebut the presumption that an investor who buys or sells stock at the market price has relied on any public material misrepresentations, at the class certification stage in a securities class action by showing a lack of stock price impact following a misrepresentation, and not following a corrective disclosure.

5 Cases that cite this headnote

**[20] Securities Regulation** 🔑 Presumptions and burden of proof

The presumption that an investor who buys or sells stock at the market price has relied on any public material misrepresentations can be rebutted by evidence that the asserted misrepresentation, or its correction, did not affect the market price of the defendant's stock, at the merits stage of a securities class action.

**[21] Federal Civil Procedure** 🔑 Common interest in subject matter, questions and relief; damages issues

In order to establish that the calculation of damages is a common question susceptible of measurement on a classwide basis, for purposes of the predominance requirement for class certification, plaintiffs must demonstrate that their theory of damages is consistent with their theory of liability. Fed. R. Civ. P. 23(b)(3).

3 Cases that cite this headnote

**[22] Federal Civil Procedure** 🔑 Stockholders, investors, and depositors

Investor's theory of damages was consistent with proposed theory of liability, so that investor established that calculation of damages was common question susceptible of measurement on classwide basis, for purposes of the predominance requirement for certification of proposed class consisting of those who purchased or otherwise acquired corporation's class A common stock within certain time period and who were damaged as a result, in putative class action against corporation and its officer alleging, inter alia, that defendants made material misrepresentations to artificially inflate corporation's stock

price in violation of § 10(b) and Rule 10b-5; investor's expert analysis clearly articulated means by which investor intended to go about calculating damages on a classwide basis, including accounting for how stock-price inflation varied during class term. Securities Exchange Act of 1934, § 10(b), [15 U.S.C.A. § 78j](#); [17 C.F.R. § 240.10b-5](#); Fed. R. Civ. P. 23(b)(3).

[2 Cases that cite this headnote](#)

### **[23] Federal Civil Procedure** [Stockholders, investors, and depositors](#)

Investor established that class certification was superior to other available methods of adjudication, for purposes of certification of proposed class consisting of those who purchased or otherwise acquired corporation's class A common stock within certain time period and who were damaged as a result, in putative class action against corporation and its officer alleging, inter alia, that defendants made material misrepresentations to artificially inflate price of corporation's stock in violation of § 10(b) and Rule 10b-5; investor's arguments included that there was no other ongoing litigation, that cost of bringing individual suits would outweigh recovery, and that action presented no likely management difficulties. Securities Exchange Act of 1934, § 10(b), [15 U.S.C.A. § 78j](#); [17 C.F.R. § 240.10b-5](#); Fed. R. Civ. P. 23(b)(3).

[1 Case that cites this headnote](#)

## **Attorneys and Law Firms**

**\*443** [Jennifer Barrett Poppe](#), Vinson & Elkins, [Alitheia Z. Sullivan](#), [Gary Ewell](#), Ewell, Brown, Blanke & Knight LLP, Austin, TX, [Michael C. Holmes](#), [Stephen S. Gilstrap](#), Vinson & Elkins LLP, Dallas, TX, for Defendants.

## **ORDER**

[SAM SPARKS](#), SENIOR UNITED STATES DISTRICT JUDGE

BE IT REMEMBERED on this day the Court considered the file in the above-styled cause, and specifically Plaintiff John Rooney's Amended Motion to Certify Class [# 113], Defendants EZCORP, Inc. (EZCORP) and Mark Kuchenrither (collectively, Defendants)'s Response [# 114] in opposition, Plaintiff's Reply [# 115] in support, and Defendants' Surreply [# 116] in opposition.

Having reviewed the documents, the governing law, the arguments of counsel, and the file as a whole, the Court now issues the following opinion and order.

## Background

This is a securities fraud class action brought on behalf of all persons who purchased Class A common stock of EZCORP—a company which provides “instant cash” services like payday loans and pawn loans—between January 28, 2014 and October 20, 2015 (the Class Period). Plaintiff alleges that during the Class Period, EZCORP CEO Mark Kuchenrither<sup>1</sup> made material misrepresentations to shareholders regarding the impact of a subsidiary's loan portfolio upon EZCORP's reported financials and thereby violated §§ 10(b) and 20(a) of the Securities Exchange Act of 1934 and SEC Rule 10b-5.

<sup>1</sup> Kuchenrither has served in several executive positions for EZCORP from the time he was hired as Senior Vice President, Strategic Development in March 2010. Second. Am. Compl. [# 47] ¶ 34. In October 2012, Kuchenrither assumed the role of CFO, and served as EZCORP's CFO until May 26, 2015. *Id.* ¶¶ 34, 149. Kuchenrither also served as EZCORP's CEO from July 18, 2014, to February 1, 2015, and as a director from August 12, 2014, to February 3, 2015. *Id.* ¶ 34. Moreover, Kuchenrither was appointed as a member of both EZCORP's and Grupo Finmart's Board of Directors during the Class Period. *Id.* ¶ 5.

### I. Alleged Accounting Failures

Between January 2012 and June 2014, EZCORP acquired a 94 percent ownership interest in Grupo Finmart. Grupo Finmart is a Mexican company which issues small consumer loans to Mexican governmental employees. The loans issued by Grupo Finmart are backed by payroll withholding agreements (“convenios”) with Mexican employers, and under these agreements, interest and principal payments are collected by the employers through payroll deductions and then remitted to Grupo Finmart. Plaintiff alleges that throughout the Class Period, EZCORP's lack of internal controls over financial reporting gave rise to two primary accounting errors \*444 in connection with Grupo Finmart's loans.

First, Plaintiff alleges EZCORP failed to properly account for Grupo Finmart's non-performing payroll loans (Non-Performing Loans). Non-Performing Loans are “loans that were being carried as active loans but with respect to which Grupo Finmart was not currently receiving payments.” Second Am. Compl. [# 47] ¶ 99. Further, there are two types of Non-Performing Loans: in-payroll loans and out-of-payroll loans. Out-of-payroll loans are outstanding loans from customers who are no longer employed. “Under Grupo Finmart's historic accounting policy,” “[i]f one payment of an out-of-payroll loan is delinquent, that one payment is considered in default; if two or more

payments are delinquent at any time, the entire loan is considered in default.” *Id.* Upon default of an out-of-payroll loan, Grupo Finmart ceased accruing future interest revenue. *Id.* However, “[d]ue to the likelihood of ultimately receiving payment if the customer remains employed, [Grupo Finmart] continue[d] to accrue interest on all in-payroll loans, even though Grupo Finmart may not be currently receiving payments.” *Id.* In its corrective disclosures, EZCORP determined Grupo Finmart's Non-Performing Loans included a number of out-of-payroll loans that had not been properly classified as such, and some in-payroll loans that had been in non-performing status for some time. *Id.* By failing to properly account for the Non-Performing Loans, Plaintiff argues, EZCORP was able “to artificially maintain its ratio of bad debt expense to consumer loan fees and interest – a measure of health of the underlying loan portfolio.” *Id.* ¶ 108.

Second, Plaintiff contends EZCORP failed to properly account for the sale of Grupo Finmart loans (Loan Sales). Between 2014 and 2015, EZCORP executed five separate sales of Grupo Finmart loans. Under the terms of the Loan Sales, third-party purchasers retained a right to return non-performing loans to EZCORP. And because the loan sales were conditional on the performance of the loans, generally accepted accounting principles (GAAP) prohibited EZCORP from recognizing any revenue from these loan sales. EZCORP disregarded this prohibition and recognized tens of millions of dollars in gains on the sales. Plaintiff claims the improper accounting for the sale of the loans had the effect of artificially boosting EZCORP's reported income in the 2014 fiscal year by 45% and its reported income during the first quarter of 2015 by 32%.

## II. Alleged False and Misleading Misstatements

The statements Plaintiff identifies as misleading are taken from EZCORP's press releases, conference calls, and SEC forms disclosing EZCORP's financial results during the Class Period. These statements deal with EZCORP's financial results during the fourth quarter of 2013 (4Q13), the 2014 fiscal year (FY2014), and the first quarter of 2015 (1Q15). In general, the statements fall into two categories (1) statements relating to the overstatement of EZCORP's financial results, as a result of EZCORP's failure to properly account for the Loan Sales and Non-Performing Loans, and (2) statements relating to the nature of the Loan Sales. According to Plaintiff, Kuchenrither knew all of the statements described above were materially false and misleading at the time they were made.

Between April 2015 and November 2015, Defendants issued a series of corrective disclosures concerning Grupo Finmart's loan portfolio and its effect upon EZCORP's reported financials. For example, on April 30, 2015, EZCORP announced the release of its 2Q15 financial results would be delayed “due to an ongoing review of certain elements of its Grupo Finmart loan portfolio, which is not yet completed.” *Id.* ¶ 96. In that same press release, EZCORP further stated it “did not undertake any asset sales in Grupo Finmart this quarter” and “noted some differences in the performance of parts of our Grupo Finmart loan portfolio that prompted a more thorough review and analysis of our loan reserves[.]” *Id.* ¶ 96. Following this announcement, EZCORP's stock fell


\$ 0.79 per share to close at \$ 8.41 per share on May 1, 2015. *Id.* ¶ 97. Further corrective disclosures also coincided with declines in the value of EZCORP's stock.

\*445 On November 9, 2015, EZCORP filed its restated financials from 2Q12 through 1Q15. The Restatement revealed, among other things, EZCORP's operating income was overstated by \$ 90.7 million, or 27.3%, during the restated periods, and its earnings per share were overstated by \$ 0.78, or 36.8%, during the restated periods. Following the filing of its restated financial results, EZCORP's stock declined \$ 0.29 per share to close at \$ 6.51 per share on November 9, 2015.

### III. Procedural Posture

Plaintiff filed this lawsuit on July 20, 2015, alleging that Defendants' false and misleading statements caused EZCORP's stock to trade at artificially inflated prices and that Plaintiff suffered financial losses following the release of EZCORP's restated financial reports. *See* Compl. [# 1]. Plaintiff now moves for class certification under [Federal Rule of Civil Procedure 23\(b\)\(3\)](#). Mot. Certify [# 113]. Plaintiff specifically seeks (1) certification of a class consisting of “all persons and entities that purchased or otherwise acquired EZCORP, Inc. Class A common stock between January 28, 2014 and October 20, 2015, inclusive, and were damaged thereby”; and (2) the appointment of Plaintiff as representative and the appointment of Block & Leviton LLP and Glancy Prongay & Murray LLP as Class Counsel and Kendall Law Group, PLLC as Liaison Counsel. Mot. Certify [# 113-1] at 7. This pending motion is ripe for review.

### Analysis


[1] Plaintiffs seeking to certify a class under [Rule 23](#) bear the burden of establishing the prerequisites to certification have been met.  *Amgen Inc. v. Conn. Ret. Plans & Tr. Funds*, 568 U.S. 455, 133 S.Ct. 1184, 1192, 185 L.Ed.2d 308 (2013). [Rule 23\(a\)](#) sets forth four such prerequisites: numerosity, commonality, typicality, and adequacy. [FED. R. CIV. P. 23\(a\)\(1\)–\(4\)](#). In addition to these baseline prerequisites, plaintiffs seeking to certify a class action under [Rule 23\(b\)\(3\)](#) must also establish that questions common to the class predominate over those questions affecting individual class members (the “predominance” requirement) and that class adjudication is superior to other available methods of fairly and efficiently adjudicating the controversy (the “superiority” requirement). [FED. R. CIV. P. 23\(b\)\(3\)](#). The Court proceeds by examining each of these requirements in turn.

#### I. Numerosity

[2] To meet the numerosity requirement, the plaintiff must establish “the class is so numerous that joinder of all members is impracticable.” [FED. R. CIV. P. 23\(a\)\(1\)](#). Here, Plaintiff seeks to certify

a class consisting of all purchasers of EZCORP's publicly traded securities during the proposed class period. Mot. Certify [# 113-1] at 7. Although the exact number of class members is unknown, EZCORP had more than 50 million shares of Class A common stock outstanding during the class period, and the average weekly trading volume on the NASDAQ Stock Market during the class period was roughly 2.7 million shares. Mot. Certify [# 113-4] Ex. A (Coffman Report) at 13. Absent any argument to the contrary, the Court concludes this evidence is sufficient to establish the proposed class is so numerous as to render joinder impracticable.<sup>2</sup>


<sup>2</sup>

Cf.  *Zeidman v. J. Ray McDermott & Co.*, 651 F.2d 1030, 1039 (5th Cir. 1981) (“The [numerosity] prerequisite ... is generally assumed to have been met in class action suits involving nationally traded securities.”)

## II. Commonality

[3] To meet the commonality requirement, the plaintiff must establish “there are questions of law or fact common to the class.” FED. R. CIV. P. 23(a)(2). In this case, Plaintiff alleges Defendants made uniform representations and omissions to the class concerning the accuracy of EZCORP's publicly reported financial results during the class period. Mot. Certify [# 113-1] at 11. These allegations implicate multiple questions common to the class, including whether Defendants' statements or omissions violated federal securities law, whether Defendants' acted with scienter, and the extent to which the market price of EZCORP's Class A shares was affected by various public statements and disclosures \*446 made by Defendants. Accordingly, the Court concludes that Plaintiff has shown there are common questions of law and fact sufficient to establish commonality.

## III. Typicality

[4] To meet the typicality requirement, the plaintiff must establish that “the claims or defenses of the representative part[y] are typical of the claims or defenses of the class.” FED. R. CIV. P. 23(a)(3). Here, Plaintiff's claims and defenses are typical of those of the class. All putative class members, including Plaintiff, allegedly purchased EZCORP stock during the class period at prices inflated by Defendants' misstatements and omissions. Mot. Certify [# 113-1] at 7, 11. Further, Plaintiff has presented a classwide legal theory and does not appear to be subject to any unique defense inapplicable to the class as a whole. See, e.g.,  *In re Schering Plough Corp. ERISA Litigation*, 589 F.3d 585, 597–99 (3d Cir. 2009) (“[The typicality inquiry] properly focuses on the similarity of the legal theory and legal claims; the similarity of the individual circumstances on which those theories and claims are based; and the extent to which the proposed representative may face significant unique or atypical defenses to her claims.”). In sum, the interests and incentives of Plaintiff and the members of the proposed class appear to be aligned, and absent any argument to



contrary from Defendants, the Court concludes that Plaintiff has established his claims are typical of those of the class.

#### IV. Adequacy



[5] [6] [7] To meet the adequacy requirement, the plaintiff must establish that he will “fairly and adequately protect the interests of the class” in his capacity as class representative. [FED. R. CIV. P. 23\(a\)\(4\)](#). The purpose of this requirement is to “uncover conflicts of interest between named parties and the class they seek to represent.” [Amchem Prods., Inc. v. Windsor](#), 521 U.S. 591, 625, 117 S.Ct. 2231, 138 L.Ed.2d 689 (1997). Moreover, in the Fifth Circuit,<sup>3</sup> the plaintiff must show that he is willing and able to “vigorously prosecute the interests of the class through qualified counsel.” [Berger v. Compaq Computer Corp.](#), 257 F.3d 475, 482–84 (5th Cir. 2001) (quoting [Gonzales v. Cassidy](#), 474 F.2d 67, 72–73 (5th Cir. 1973) ). And although class representatives “need not be legal scholars and are entitled to rely on counsel,” the plaintiff must at least “know more than that they were involved in a bad business deal.” [Berger](#), 257 F.3d at 483.



3

Compare [Horton v. Goose Creek Indep. Sch. Dist.](#), 690 F.2d 470, 484–85 (5th Cir. 1982) (“The adequacy requirement mandates an inquiry into the zeal and competence of the representative's counsel and into the willingness and ability of the representative to take an active role in and control the litigation and to protect the interests of absentees.” (relying on [Jaurigui v. Ariz. Bd. of Regents](#), 82 F.R.D. 64 (D. Ariz. 1979); [Klein v. Miller](#), 82 F.R.D. 6, 8 (N.D. Tex. 1978) )), with [In re Literary Works in Elec. Databases Copyright Litig.](#), 654 F.3d 242, 249 (2d Cir. 2011) (“Adequacy is twofold: the proposed class representative must have an interest in vigorously pursuing the claims of the class, and must have no interests antagonistic to the interests of other class members.”), [In re Gen. Motors Corp. Pick-Up Truck Fuel Tank Prods. Liab. Litig.](#), 55 F.3d 768, 800 (3d Cir. 1995) (“The adequacy of representation inquiry ... considers whether the named plaintiffs' interests are sufficiently aligned with the absentees', and it tests the qualifications of counsel to represent the class.”), and [Ellis v. Costco Wholesale Corp.](#), 657 F.3d 970, 985 (9th Cir. 2011) (“To determine whether plaintiffs will adequately represent a class, ... [courts must ask whether] named plaintiffs and their counsel have any conflicts of interests with other class members and ... [whether] the named plaintiffs and their counsel [will] prosecute the action vigorously on behalf of the class[.]” (internal quotation marks and citation omitted) ).


[8] As an initial matter, Defendants argue that Plaintiff is not an adequate class representative because, as an individual with a small amount of purported damages, Plaintiff is not the PSLRA's preferred type of class representative. Resp. [# 114] at 17; *see also id.* at 11 (arguing the adequacy inquiry must be “particularly searching” in securities class actions implicating the PSLRA). This argument misses the mark because the PSLRA's provisions governing the selection of class





representatives do not affect the inquiry into whether a proposed class representative has met Rule 23 (a)(4)'s adequacy requirement. See  *Berger v. Compaq Computer Corp.*, 279 F.3d 313, 313–14 (5th Cir. 2002) (per curiam) (noting the PSLRA did not “change the law ... regarding the standard for conducting a rule 23(a)(4) adequacy inquiry”); \*447 see also  *In re Cavanaugh*, 306 F.3d 726, 736 (9th Cir. 2002) (“[T]he [PSLRA] did not change the standard for adequacy[ ] ....”).<sup>4</sup>

<sup>4</sup> But see  *In re Kosmos Energy Ltd. Sec. Litig.*, 299 F.R.D. 133, 146 (N.D. Tex. 2014) (interpreting the original panel decision in  *Berger* to require a “particularly searching” investigation into adequacy of the proposed class representative”).

[9] Defendants also suggest Plaintiff is an inadequate class representative because he lives in Dublin, Ireland. Resp. [# 114] at 16–17. Defendants do not, however, clearly explain why or how living in Dublin might affect Plaintiff's ability to serve as a class representative. See Resp. [# 114] at 16 (stating only that Plaintiff's residence in Dublin is a “relevant factor[ ]” that “undermine[s] his ability to protect the interests of absent class members”). And in any event, the Court concludes Plaintiff's residence in Ireland does not affect his ability to protect the interests of the class or his ability to prosecute this action through qualified counsel. Plaintiff can easily correspond with counsel through email, telephone, and video chat to stay abreast of developments in the case. There is also little reason to believe Plaintiff's residence in Ireland poses an obstacle to his ability to appear in court as needed. Not only has Plaintiff declared he will be available to travel to trial in Austin as necessary, but Plaintiff has already flown to Boston to sit for a deposition on February 9, 2018. Rooney Decl. at 6; Mot. Certify [# 113-1] at 13. The Court therefore concludes that Plaintiff's Irish residency would not render Plaintiff an inadequate class representative.


[10] Finally, Defendants contend Plaintiff is not an adequate class representative because Plaintiff has no idea what is going on in this lawsuit. Resp. [# 114] at 15. To be sure, Plaintiff could have a better grasp on the specifics of this lawsuit, see Resp. [# 114] at 15 (detailing the things that Plaintiff did not know at his deposition), and he is not the Platonic ideal of a class representative. Yet he has also demonstrated that he knows more than that he was involved in a bad business deal.  *Berger*, 257 F.3d at 483. Plaintiff testified at his deposition that he bought shares of EZCORP on the basis of its past performance, that the value of his EZCORP stock dropped precipitously after he purchased it, and that he attributes this drop in value to EZCORP's restatements of its financials. Mot. Certify Class [# 113-8] Ex. A at 10 (“[I]t wasn't just, you know, ... that they hoped to get a new customer and they just missed out.... Those things I understand. But just having to restate accounts, it just seemed, I just felt that I was taken in and then ended up losing[ ] ....”). Plaintiff also testified at his deposition that he believed EZCORP violated federal securities law by misleadingly classifying some of the company's loans in order to report a profit instead of a loss. *Id.* at 11–13. In light of this testimony and additional declarations made by Plaintiff in conjunction with his motion for class certification, see Rooney Decl. at 1–7, the Court concludes Plaintiff has

demonstrated sufficient knowledge regarding the subject matter of this litigation to serve as an adequate class representative.

[11] Having dispensed with Defendants' objections, the Court concludes Plaintiff has established that he will fairly and adequately protect the interests of the class. The Court is aware of no conflicts between Plaintiff and the members of the proposed class, and as best the Court can tell, Plaintiff's interests are aligned with those of the class as a whole.  *Amchem*, 521 U.S. at 625, 117 S.Ct. 2231. Moreover, Plaintiff has demonstrated he is both willing and able to vigorously prosecute<sup>5</sup> the interests of the class through qualified counsel.  *Gonzales*, 474 F.2d at 72–73. In brief, the Court finds Plaintiff has met the adequacy requirement.



<sup>5</sup> Indeed, this lawsuit has been vigorously contested by both sides for the better part of two years.



## V. Predominance

[12] To meet the predominance requirement, the plaintiff must establish that “questions of law or fact common to class members predominate over any questions affecting only individual members.” FED. R. CIV. P. 23(b)(3). This inquiry tests whether the proposed class is “sufficiently cohesive to warrant \*448 adjudication by representation.”  *Amchem*, 521 U.S. at 623, 117 S.Ct. 2231.

Defendants suggest two ways in which Plaintiff has failed to establish predominance. First, Defendants argue that some class members will have to demonstrate reliance on an individual basis and that these individual inquiries will predominate over common questions as to those class members. Resp. [# 114] at 19–26. Second, Defendants argue that Plaintiff cannot demonstrate predominance because Plaintiff has not “articulate[d] how alleged damages could be calculated on a class-wide basis.” Resp. [# 114] at 17. The Court evaluates both arguments and then turns to the ultimate question of whether Plaintiff has met the predominance requirement.



### A. Reliance



[13] [14] In securities law cases, predominance often hinges upon whether or not class members will need to individually demonstrate they relied upon a misrepresentation or omission made by the defendant. See  *Amgen Inc. v. Conn. Ret. Plans & Tr. Funds*, 568 U.S. 455, 133 S.Ct. 1184, 1193–96, 185 L.Ed.2d 308 (2013) (noting plaintiffs seeking to recover damages under § 10(b) must prove they relied on such a misrepresentation or omission when purchasing or selling a security). If the securities at issue traded in an efficient market, the class members may sometimes invoke a rebuttable presumption of reliance. See  *Halliburton Co. v. Erica P. John Fund, Inc.* (“*Halliburton II*”), 573 U.S. 258, 134 S.Ct. 2398, 2408, 189 L.Ed.2d 339 (2014)

(“[W]henver the investor buys or sells stock at the market price, his ‘reliance on any public material misrepresentations ... may be presumed for purposes of a Rule 10b-5 action.’” (alteration in original) (quoting  *Basic Inc. v. Levinson*, 485 U.S. 224, 247, 108 S.Ct. 978, 99 L.Ed.2d 194 (1988) ) ). This presumption is of particular use in class actions because to the extent the class members are able to invoke the presumption, reliance becomes a common question capable of resolution on a classwide basis. Conversely, if the class members cannot establish they are entitled to rely on the presumption, or if the defendant successfully rebuts it, class members must individually establish reliance, and in practice, such individual inquiries almost inevitably prevent the plaintiff from establishing predominance by overwhelming whatever common questions may exist. See  *Unger v. Amedisys Inc.*, 401 F.3d 316, 322 (5th Cir. 2005) (“Absent an efficient market, individual reliance by each plaintiff must be proven, and the proposed class will fail the predominance requirement.”)).


The Court proceeds by first assessing whether Plaintiff is entitled to rely on the presumption. It then considers whether Defendant has rebutted the presumption.



### 1. The *Basic* Presumption

[15] To establish the class is entitled to rely on the  *Basic* presumption, the plaintiff must demonstrate (1) “the alleged misrepresentations were publicly known”; (2) the alleged misrepresentations were material; (3) “the stock traded in an efficient market”; and (4) putative class members “traded the stock between the time the misrepresentations were made and when the truth was revealed.”  *Halliburton II*, 134 S.Ct. at 2408.






[16] Here, Plaintiff has met all four prerequisites to invoking the presumption at the class certification stage.<sup>6</sup> First, the alleged misrepresentations were publically known. As detailed in the Court's previous orders, the alleged misrepresentations and corrective statements made by Defendants were contained within EZCORP's press releases, SEC forms, and conference calls with analysts and investors. See Order of July 26, 2018 [# 102] at 4. Second, the alleged misrepresentations are presumed to be material at the class certification stage. See  *Amgen*, 133 S.Ct. at 1191, 1195-96 (“Proof [of materiality] is not a prerequisite to class certification.”). Third, EZCORP's stock traded in an efficient market. EZCORP's stock traded in the NASDAQ Stock Market, one of the largest markets in the world, and perhaps for this reason, Defendants have not contested that \*449 EZCORP stock trades in an efficient market. Hr'g Tr. [# 119] at 4. Nevertheless, the Court has reviewed the expert report submitted by Plaintiff in support, see Coffman Report at 12-34, and applying the  *Cammer* factors,<sup>7</sup> the Court concludes Plaintiff has established EZCORP's stock traded in an efficient market. Fourth and finally, Plaintiff and the




rest of the proposed class traded the stock between the time the misrepresentations were made and when they were corrected. *See* Third Am. Compl. [# 106] at 28 (specifying that first alleged misrepresentation was contained within press release issued by EZCORP on January 28, 2014); Order of July 26, 2018 [# 102] (outlining timing of corrective disclosures); Mot. Certify [# 113] at 2 (defining class as “all persons and entities that purchased or otherwise acquired [EZCORP stock] between January 28, 2014 and October 20, 2015). In sum, the Court concludes Plaintiff has established the class is entitled to a presumption of reliance at the class certification stage.







6 Indeed, Defendants do not contend that Plaintiff has failed to establish any of the prerequisites to invoking the  *Basic* presumption. *See* Resp. [# 114] at 17–26 (seeking to rebut presumption but declining to contest Plaintiff’s ability to invoke the presumption in the first place).








7 Though neither the Supreme Court nor the Fifth Circuit has adopted a formal test for market efficiency, district courts routinely apply a list of factors derived from  *Cammie v. Bloom*, 711 F.Supp. 1264 (D.N.J. 1989). *See, e.g.,*  *In re Petrobras Sec. Litig.*, 3121 F.R.D. 354, 364–65 (S.D.N.Y. 2016).


## 2. Rebutting the *Basic* Presumption


[17] [18] The  *Basic* presumption relies upon indirect evidence of price impact to establish reliance; direct evidence that the misrepresentation did not affect the price of the stock severs that link.  *Halliburton II*, 134 S.Ct. at 2414. Thus, in  *Halliburton II*, the Supreme Court held that defendants can defeat the  *Basic* presumption at the class certification stage “through evidence that the misrepresentation did not in fact affect the stock price.”  *Id.* Here, Defendants argue they have partially rebutted the presumption by showing there was no statistically significant price adjustment following two of the corrective disclosure dates identified by Plaintiff. *See* Resp. [# 114] at 19–26 (asserting there was no statistically significant price adjustment following disclosures on July 17, 2015 and October 20, 2015). As explained by Defendants, if there was no statistically significant price adjustment following the disclosures, any price adjustment that did occur must be due to “random chance,” and thus, the misrepresentation could not have affected the stock price. *Id.* at 22. This line of reasoning suffers from several flaws.

[19] [20] As an initial matter, Defendant have failed to rebut the  *Basic* presumption because Defendants have only pointed to evidence that there was no statistically significant price adjustment following two of the *corrective disclosure* dates. Resp. [# 114] at 21. By contrast,  *Halliburton II* allows defendants to defeat the  *Basic* presumption “through evidence that the

*misrepresentation* did not in fact affect the stock price.”  134 S.Ct. at 2414. Defendants argue they should be able to rebut the presumption by showing the absence of a statistically significant price adjustment following either the misrepresentation *or* the corrective disclosure. Resp. [# 114] at 22–23. But the Court concludes  *Halliburton II* only allows defendants to rebut the  *Basic* presumption by showing a lack of price impact following a misrepresentation, and not following a corrective disclosure. Compare  *Halliburton II*, 134 S.Ct. at 2414, 2416 (holding defendants may rebut  *Basic* presumption by showing lack of “price impact”), with  *Erica P. John Fund, Inc. v. Halliburton Co.* (“*Halliburton I*”), 563 U.S. 804, 814, 131 S.Ct. 2179, 180 L.Ed.2d 24 (2011) (defining “price impact” as “the effect of a misrepresentation on a stock price” and holding that the question of whether a misrepresentation “caused a subsequent economic loss” when corrected is a matter of loss causation that need not be established by plaintiff at the class certification stage).<sup>8</sup>

<sup>8</sup> Defendants protest that  *Halliburton II* “made clear” the  *Basic* presumption could be rebutted “ ‘by evidence that the asserted misrepresentation (*or its correction* ) did not affect the market price of the defendant's stock.’ ” Resp. [# 114] at 22–23 (quoting  *Halliburton II*, 134 S.Ct. at 2414) (emphasis added) ). This is undoubtedly true—at the merits stage. But this case is currently at the class certification stage, and because the language from  *Halliburton II* relied upon by Defendants is quite clearly discussing the means by which Defendants can rebut the  *Basic* presumption *at the merits stage*, it has no bearing here.  *Halliburton II*, 134 S.Ct. at 2414 (“There is no dispute that defendants may introduce such evidence at the merits stage to rebut the  *Basic* presumption[ ] ... including evidence that the asserted misrepresentation (*or its correction*) did not affect the market price of the defendant's stock.”).





\*450 Defendants' attempt to rebut the  *Basic* presumption is also flawed from a statistical perspective. Defendants suggest the lack of a statistically significant price adjustment following a corrective disclosure shows that whatever price adjustment has occurred must be due to “random chance” rather than a predicate misrepresentation. Resp. [# 114] at 22. But that is not how hypothesis testing works. A statistically significant price adjustment following a corrective disclosure is evidence the original misrepresentation did, in fact, affect the stock price. The converse, however, is not true—the absence of a statistically significant price adjustment does *not* show the stock price was unaffected by the misrepresentation.<sup>9</sup> Nor does it indicate that what price adjustment did occur must be attributed to “random chance.” Resp. [# 114] at 22.


<sup>9</sup> See generally  *In re Petrobras Sec.*, 862 F.3d 250, 278–79 & n.30 (2d Cir. 2016) (observing that “ ‘the failure of the price to react so extremely as to be detectable ... mean[s] only that’ the effect size was not large enough to be detected in the available sample” and that “ ‘[w]hile






some courts have been sensitive to this distinction ..., other courts have remained inattentive to this fact, which has generated inaccurate findings in some securities cases.’ ” (quoting Alon Brav & J.B. Heaton, *Event Studies in Securities Litigation: Low Power, Confounding Effects, and Bias*, 93 WASH. U. L. REV. 583, 602 (2015) ).

Here, Plaintiff's expert, Chad Coffman, submitted an expert report indicating the two corrective disclosure dates at issue here returned p-values of 0.234 and 0.233, respectively. Resp. [# 114-3] Ex. 2 at 9–11. These p-values suggest there is a 77 percent chance the corrective disclosures identified by Plaintiff negatively impacted EZCORP's stock price on these dates. *See id.* What they do not suggest is that the misrepresentation “did not affect the stock price.”<sup>10</sup>

<sup>10</sup> Defendants suggest that even though there is a 77 percent chance the corrective disclosures negatively impacted EZCORP'S stock price, they should nevertheless be allowed to rebut the  *Basic* presumption by pointing to the absence of a statistically significant price impact at a 95-percent confidence interval. Resp. [# 114] at 21–22. But the practical effect of such a maneuver would be to require plaintiffs to show loss causation at the class certification stage, and  *Halliburton I* held that plaintiffs need only make such a showing *after* class certification. *See*  *Halliburton I*, 563 U.S. at 813, 131 S.Ct. 2179 (“Loss causation ... requires a plaintiff to show that a misrepresentation that affected the integrity of the market price *also* caused a subsequent economic loss[ ] ... [and] has nothing to do with whether an investor relied on the misrepresentation in the first place.”); *cf.*  *In re Petrobras Sec.*, 862 F.3d at 278–279 (suggesting district courts should not rely solely upon directional event studies at the class certification stage because “methodological constraints limit their utility in the context of single-firm analyses”).

In brief, the Court concludes that Defendants have failed to rebut the  *Basic* presumption because they have not pointed to any evidence that the alleged misrepresentations did not affect the stock price. As a result, Plaintiff is entitled to rely on the presumption at the class certification stage. And insofar as the class can rely on the presumption to establish reliance, the class's ability to demonstrate reliance remains a common question capable of classwide resolution.

## B. Damages Calculations

[21] In order to establish that the calculation of damages is a common question “susceptible of measurement” on a classwide basis, plaintiffs must demonstrate that their theory of damages is consistent with their theory of liability. *See*  *Ludlow v. BP, PLC*, 800 F.3d 674, 688–89 (5th Cir. 2015) (interpreting  *Comcast Corp. v. Behrend*, 569 U.S. 27, 35, 133 S.Ct. 1426, 185 L.Ed.2d 515 (2013) ), cert denied, — U.S. —, 136 S.Ct. 1824, 194 L.Ed.2d 829 (2016); *see also*  *Comcast*, 569 U.S. at 35, 133 S.Ct. 1426 (“[A] model purporting to serve as evidence of damages” suffered


by the class “must measure only those damages attributable to that theory [of liability relied on by the plaintiff].”).

[22] Defendants argue that Plaintiff’s proposed theory of damages “fails to articulate how alleged damages could be calculated on a class-wide basis” consistent with the proposed theory of liability. Resp. [# 114] at \*451 17. Specifically, Defendants contend that Plaintiff’s expert, Chad Coffman, has “failed to articulate a means to”: (1) “disaggregate from damages price declines resulting from non-actionable confounding information”; (2) “isolate the impact of materialization of known risks from the impact of allegedly concealed risks”; (3) “calculate damages on a class-wide basis if Plaintiff is able to prevail only on claims relating to certain of the remaining alleged misstatements”; and (4) “account for how stock-price inflation varied during the purported class term.” Resp. [# 114] at 19.

In fact, Coffman’s expert report considered all of these things. In his report, Coffman explains that he conducted an event study and used regression analysis to assess the effect of EZCORP’s disclosures upon the company’s stock. Coffman Report at 22–23. Specifically, for each trading day analyzed, Coffman constructed a regression model using data from the prior 120 trading days and leveraged this “‘rolling’ estimation window” to discern the relationship between EZCORP’s common stock, industry and market factors, and firm-specific price volatility. *Id.* at 23–30. Using this methodology, Coffman calculated the abnormal returns attributable to “new material news and changes in the market price of EZCORP Common Stock.” *Id.* at 30. Coffman explains in his report that these abnormal returns can be used to calculate damages on a classwide basis using the “out-of-pocket” method, “which measures damages as the artificial inflation per share at the time of purchase less the artificial inflation at the time of sale.” Coffman Report at 34–35. In this way, damages for individual class members can “be calculated formulaically based on information collected in the claims process ....” *Id.*

As far as the Court can discern, this analysis quite clearly articulates the means by which Coffman intends to go about calculating damages on a classwide basis. And given Defendants’ complete failure to expand on any of their extremely conclusory complaints regarding Coffman’s report, the Court concludes that Plaintiff’s theory of damages is consistent with the proposed theory of liability and that Plaintiff has established the calculation of damages is a common question susceptible of measurement on a classwide basis.

### C. Conclusion

The Court concludes this lawsuit implicates a host of common questions, including: whether the class may invoke the  *Basic* presumption at the merits stage to establish reliance; the calculation of damages; whether Defendants’ statements or omissions violated federal securities law; whether Defendants acted with scienter; and the extent to which the market price of EZCORP’s Class A



shares was affected by various public statements and disclosures made by Defendants. In light of these common questions identified by Plaintiff, the Court finds that Plaintiff has established common questions predominate over questions affecting only individual members. Accordingly, Plaintiff has met the predominance requirement.

## **VI. Superiority**

[23] To establish superiority, the plaintiff must demonstrate that class certification is “superior to other available methods for fairly and efficiently adjudicating the controversy.” [FED. R. CIV. P. 23\(b\)\(3\)](#). [Rule 23](#) sets forth several “matters pertinent” to a finding of superiority, including: (1) the class members' interests in individually controlling the prosecution or defense of separate actions; (2) the extent and nature of any litigation concerning the controversy already begun by or against class members; (3) the desirability or undesirability of concentrating the litigation of the claims in the particular forum; and (4) the likely difficulties in managing a class action. [FED. R. CIV. P. 23\(b\)\(3\)\(A\)–\(D\)](#).

Here, Plaintiff argues that there is no other ongoing litigation concerning this controversy; that the individual class members have little interest in controlling the prosecution because the cost of bringing individual suits to seek recovery would in most cases outweigh the recovery obtained; that this forum is as desirable a forum as any given the geographic dispersal of investors and EZCORP's headquarters in Austin, Texas; and that this putative class action presents \*452 no likely management difficulties. Mot. Certify [# 113-1] at 25–26. Defendants have not put forward any argument in response, and the Court concludes that Plaintiff has established that class certification is superior to other available methods of adjudication.

## **Conclusion**

The Court concludes that Plaintiff's motion for class certification should be granted because Plaintiff has met all predicate requirements for bringing a class action under [Rule 23\(b\)\(3\)](#).

Accordingly,

IT IS ORDERED that Plaintiff's Motion for Class Certification [# 113] is GRANTED.

IT IS FURTHER ORDERED that the Court CERTIFIES a class consisting of “all persons and entities that purchased or otherwise acquired EZCORP, Inc. Class A common stock between January 28, 2014 and October 20, 2015, inclusive, and were damaged thereby. Excluded from the Class are Defendants, the officers and directors of the Company, at all relevant times, members of their immediate families and their legal representatives, heirs, successors or assigns and any entity in which Defendants have or had a controlling interest.”

IT IS FURTHER ORDERED that the Court APPOINTS Lead Plaintiff John Rooney as Class Representative.

IT IS FINALLY ORDERED that the Court APPOINTS the law firms of Block & Leviton LLP and Glancy Prongay & Murray LLP as Class Counsel and The Kendall Law Group, PLLC as Liaison Counsel for the Class.

## **All Citations**

330 F.R.D. 439

---

End of Document

© 2023 Thomson Reuters. No claim to original U.S. Government Works.

# Exhibit 69

**IN THE UNITED STATES DISTRICT COURT  
FOR THE NORTHERN DISTRICT OF TEXAS  
DALLAS DIVISION**

PEDRO RAMIREZ, JR., Individually and  
on Behalf of All Others Similarly Situated,

Plaintiff,

v.

Civil Action No. 3:16-CV-03111-K

EXXON MOBIL CORPORATION, REX  
W. TILLERSON, ANDREW P. SWIGER,  
JEFFREY J. WOODBURY, and DAVID S.  
ROSENTHAL,

Defendants.

**MEMORANDUM OPINION AND ORDER**

**TABLE OF CONTENTS**

1. INTRODUCTION.....	3
2. BACKGROUND .....	4
2.1. Exxon Mobil’s Canadian Bitumen Operations, Including Kearl.....	5
2.2. Exxon Mobil’s RMDG Operations .....	10
2.3. Proxy Costs of Carbon .....	11
2.4. The Alleged Corrective Disclosures .....	12
2.4.1. The First Corrective Disclosure: The Guardian Article (November 9, 2015) .....	12
2.4.2. The Second Corrective Disclosure: <i>The Los Angeles Times</i> Article (January 20, 2016) .....	13
2.4.3. The Third Corrective Disclosure: Second Quarter Earnings Announcement for 2016 (July 29, 2016) .....	13
2.4.4. The Fourth Corrective Disclosure: <i>The Washington Post</i> Op-Ed (August 10, 2016) .....	13
2.4.5. The Fifth Corrective Disclosure: Third Quarter Earnings (October 28, 2016) .....	14

2.4.6.	The Sixth Corrective Disclosure: UBS Downgrade (January 18, 2017) .....	15
2.4.7.	The Seventh Corrective Disclosure: Fourth Quarter Earnings Announcement for 2016 (January 31, 2017) .....	15
3.	LEGAL STANDARDS .....	15
4.	ANALYSIS .....	16
4.1.	Rule 23(a).....	16
4.1.1.	Numerosity .....	16
4.1.2.	Commonality .....	17
4.1.3.	Typicality.....	18
4.1.4.	Adequacy .....	19
4.2.	Rule 23(b)(3) .....	22
4.2.1.	Predominance .....	22
4.2.1.1.	Fraud-on-the-Market Theory and the Presumption of Reliance .....	25
4.2.1.1.1.	Rebutting Basic’s Presumption of Reliance: Price Impact.....	28
4.2.1.1.2.	The Parties’ Event Studies .....	30
4.2.1.1.3.	Event Study Windows .....	34
4.2.1.1.4.	The Corrective Disclosures Analyzed .....	37
4.2.1.1.4.1.	November 9, 2015: <i>The Guardian</i> Article .....	38
4.2.1.1.4.2.	January 20, 2016: <i>The Los Angeles Times</i> Article.....	39
4.2.1.1.4.3.	July 29, 2016: Second Quarter Earnings Announcement for 2016.....	41
4.2.1.1.4.4.	August 10, 2016: <i>The Washington Post</i> Op-Ed .....	44
4.2.1.1.4.5.	October 28, 2016: Third Quarter Earnings Announcement for 2016.....	45
4.2.1.1.4.6.	January 18, 2017: UBS Downgrade .....	50
4.2.1.1.4.7.	January 31, 2017: Fourth Quarter Earnings Announcement for 2016.....	51
4.2.1.2.	Omissions-Based Presumption of Reliance.....	53
4.2.2.	Superiority.....	54
5.	CLASS DEFINITION AND CONCLUSION .....	55

Before the Court are Lead Plaintiff's Motion for Class Certification, Doc. No. 86, Lead Plaintiff's Memorandum of Law in Support of Motion for Class Certification (collectively, the "Motion" or the "Motion for Class Certification"), Doc. No. 87, Defendants' Corrected Memorandum of Law in Opposition to Lead Plaintiff's Motion for Class Certification (the "Response"), Doc. No. 115, and Lead Plaintiff's Reply in Further Support of Its Motion for Class Certification (the "Reply"), Doc. No. 104. Having carefully considered the Motion, the Response, the Reply, the Consolidated Complaint for Violations of the Federal Securities Laws (the "Complaint"), Doc. No. 36, and the applicable law, the Court **GRANTS in part** and **DENIES in part** the Motion.

## 1. INTRODUCTION

Lead Plaintiff Greater Pennsylvania Carpenters Fund (the "Fund" or "Plaintiff") brings this putative class action against Defendants Exxon Mobil Corporation ("Exxon Mobil" or the "Company"), Rex. W. Tillerson ("Tillerson"), Andrew P. Swiger ("Swiger"), Jeffrey J. Woodbury ("Woodbury"), and David S. Rosenthal ("Rosenthal") (collectively, "Defendants"), alleging violations of § 10(b) of the Securities and Exchange Act of 1934 (the "Exchange Act"), 15 U.S.C. § 78(j)(b), and Securities and Exchange Commission ("SEC") Rule 10(b)-5 promulgated thereunder, 17 C.F.R. § 240.10b-5. Doc. No. 36. In its Complaint, Plaintiff alleges that Defendants violated § 10(b) of Exchange Act and SEC Rule 10(b)-5 by, among other things: (1) misleading investors about Exxon Mobil's investment and asset valuation processes for proxy costs

of carbon; (2) failing to properly account for and disclose losses and requisite reserve revisions related to Exxon Mobil's Kearl Lake Operations ("Kearl" or the "Kearl Operation(s)"); and (3) failing to properly account for and disclose losses and take impairments related to Exxon Mobil's Rocky Mountain Dry Gas Operations ("RMDG" or the "RMDG Operation(s)"). *E.g.*, Doc. No. 36 at 174-78. Plaintiff also claims via § 20(a) of the Exchange Act, 15 U.S.C. § 78(t)(a), that Tillerson, Swiger, Woodbury, and Rosenthal (the "Individual Defendants") are liable as control persons of Exxon Mobil. *E.g.*, Doc. No. 36 at 178-79.

Plaintiff moves the Court to certify the following class:

All persons who purchased or otherwise acquired Exxon Mobil Corporation common stock between March 31, 2014 and January 30, 2017, inclusive, and were damaged thereby.

*E.g.*, Doc. No. 87 at 8. Plaintiff seeks appointment as class representative, and asks the Court to appoint lead counsel in this case, Robbins Geller Rudman & Dowd LLP ("Robbins Geller"), as class counsel. Doc. No. 86; Doc. No. 86-1; Doc. No. 87.

## **2. BACKGROUND**

Exxon Mobil is a multinational oil and gas company whose stock trades on the New York Stock Exchange ("NYSE") under the ticker "XOM." Doc. No. 36 ¶ 34. Plaintiff is a pension fund based in Pittsburgh, Pennsylvania. *Id.* ¶ 33. Plaintiff alleges that Defendants have violated the securities laws by misrepresenting Exxon Mobil's operations and financial health. Defendants' alleged misrepresentations fall into roughly three categories: (1) misrepresentations about Exxon Mobil's Canadian



Bitumen Operations, including Kearl; (2) misrepresentations about Exxon Mobil's RMDG Operations; and (3) misrepresentations about Exxon Mobil's proxy costs of carbon. *See* Doc. No. 87 at 10-11; Doc. No. 104 at 11. The Court briefly discusses each category and the corresponding misstatements as alleged in the Complaint below.

### **2.1. Exxon Mobil's Canadian Bitumen Operations, Including Kearl**

Exxon Mobil controls two separate upstream bitumen operations in Alberta, Canada: the Kearl Operation and the Cold Lake Operation (collectively, the "Canadian Bitumen Operations"). Doc. No. 36 ¶ 96. Kearl is a joint venture between Exxon Mobil's majority-owned and fully consolidated subsidiary, Imperial Oil Limited ("Imperial"), and Exxon Mobil's wholly-owned subsidiary, ExxonMobil Canada. *Id.* Imperial owns a 70.96% stake in Kearl, with the remaining 29.04% being held by ExxonMobil Canada. *Id.* ¶ 100. The Cold Lake Operation is entirely owned by Imperial. *Id.* ¶ 96.

In its bitumen operations, Exxon Mobil extracts raw oil from bitumen, a thick, tar-like substance found in loose sand and clay deposits. *E.g., id.* ¶¶ 92, 96, 100. The process is complex and costly. *E.g., id.* ¶¶ 40, 44. Exxon Mobil's multibillion-dollar Kearl Operation is a prime example, comprising four massive open-pit mines where Exxon Mobil and Imperial strip-mine bitumen from the earth's surface. *E.g., id.* ¶¶ 92, 96, 100. Unlike conventional light crude oil that can be easily pumped from the ground and refined, bitumen requires extensive processing before it can be transformed into useable fuel. *E.g., id.* ¶¶ 92, 95. As a result of these higher production costs, among

other factors, extracting oil from bitumen offers lower profit margins than pumping light crude oil. *E.g., id.* ¶ 95.

The success of bitumen projects like the Kearl Operation depends on the market price for oil. *See id.* ¶¶ 2, 56, 156. If oil prices drop too low, the costs associated with developing and running a project may exceed revenues. *See id.* ¶¶ 56, 63.

Investors and regulators care about whether oil and gas companies record their costs accurately. *See id.* ¶ 73. Regulators require oil and gas operators, including Exxon Mobil, to capitalize a considerable portion of the costs related to acquiring, exploring, and developing oil and gas projects. *Id.* ¶¶ 45, 55. A capitalized cost is an expense that a company treats as a long-term asset investment rather than an immediate expense; it is recognized on the balance sheet as an asset and is gradually expensed over its useful life through depreciation or amortization. *Id.* ¶¶ 45, 325. An asset's "carrying value" refers to the cost of the asset less depreciation or amortization. *Id.* ¶ 325. A capitalized asset is expected to generate future cash flows in excess of its carrying value. *See id.* ¶ 55. If circumstances change and future cash flows are no longer projected to cover the carrying value of the asset, the asset is "impaired." *Id.* ¶¶ 55, 328-29. The owner of an impaired asset must adjust its recorded value to equal its fair value by recording an impairment charge on the owner's earnings statement. *Id.* ¶ 55.

Regulators also require oil and gas companies to disclose detailed information about their "reserves." *Id.* ¶ 50. Reserves represent the quantity of raw oil and gas that a company either owns or holds the rights to extract. *Id.* ¶¶ 46, 72. They are the core

assets of an oil and gas company. *Id.* One special class of reserves are “proved reserves,” which encompass raw oil and gas that a company can profitably extract under the economic conditions existing at the time the business records the reserves in a financial statement. *Id.* ¶¶ 52, 332, 362. Companies assess profitability using historical prices—generally, the average of the first-day-of-the-month prices for the twelve months prior to the assessment—and current costs. *Id.* ¶¶ 52, 333. When later assessments reveal that reserves previously classified as proved are no longer profitable to extract, these reserves must be “de-booked,” or reclassified as unproved reserves. *Id.* ¶¶ 54, 334. Companies disclose the reserves they have classified as proved in their financial statements. *Id.* ¶ 47.

By the time Exxon Mobil had finished construction, started production, and initiated a second phase of expansion at Kearl in 2013, the Western Canadian Select benchmark (“WCS”) for heavy crude produced from the Canadian oil sands had maintained an average price of around \$72 per barrel over three consecutive years. *Id.* ¶¶ 58, 107. In 2014, however, oil and gas prices began a marked and prolonged global collapse. *Id.* ¶¶ 148-54. From its peak in June 2014, the WCS benchmark experienced an 83% decline, reaching \$14.50 per barrel in January 2016. *Id.* ¶ 148. Through 2014 and 2015, Exxon Mobil’s competitors recognized billions of dollars in asset impairments, including for Canadian oil sands projects like Kearl. *Id.* ¶¶ 156-62, 167. Exxon Mobil did not. *Id.* ¶¶ 166-68. Plaintiff alleges that Tillerson—then Exxon Mobil’s Chief Executive Officer and Chairman—and Woodbury—then Exxon Mobil’s

Secretary and Vice President of Investor Relations—instead made misleading statements to the market in an effort to portray Exxon Mobil as immune from the market forces impacting its peers. *Id.* ¶¶ 35, 37, 71, 81-84, 90, 257, 265-68, 272-75, 277-87, 289-92, 298-301, 302, 309-17.

According to Plaintiff, Defendants also misrepresented Exxon Mobil’s proved reserves in the Company’s SEC filings. On February 24, 2016, Exxon Mobil filed its 2015 Form 10-K (the “2015 10-K”) with the SEC. *Id.* ¶¶ 16, 277, 343, 348. At the time of the filing, proved reserves from the Canadian Bitumen Operations, most of which were attributable to Kearl, constituted a significant portion of Exxon Mobil’s total worldwide proved reserves, representing 31% of the company’s liquids proved reserves and 18% of the company’s combined liquids and natural gas proved reserves. *Id.* ¶¶ 97, 101, 346. Plaintiff alleges that the 2015 10-K implied that the average profit per barrel of bitumen produced from its Canadian Bitumen Operations in 2015 was \$5.87. *Id.* ¶ 343. Plaintiff contends that Exxon Mobil’s Canadian Bitumen Operations were losing money by mid-November 2015, if not sooner, and there was no indication of a favorable change in the foreseeable future. *Id.* ¶¶ 170, 301, 342, 344. While Plaintiff does not allege that the average profit implied by 2015 10-K was inaccurate (it apparently accounted for the recent losses), Plaintiff maintains that Defendants’ failure to disclose the Canadian Bitumen Operations’ losses violated Generally Accepted Accounting Principles (“GAAP”) and made the implied average profit materially misleading. *See id.* ¶¶ 17, 255-56, 287, 318.

The continued drop in oil prices also threatened the profitability of the Kearl bitumen reserves. *See id.* ¶¶ 20, 22, 107, 169, 177, 180-81, 231. Plaintiff alleges that, by the end of 2015, the Kearl reserves were—at best—on the verge of no longer meeting the SEC’s definition of proved reserves. *Id.* ¶¶ 175-76, 301. When Exxon Mobil filed the 2015 10-K in late February and oil prices still had not improved, Plaintiff claims that it was “all but certain” that the reserves would no longer meet the SEC’s definition of proved reserves by the end of 2016 and would need to be de-booked. *Id.* ¶¶ 14, 20, 169, 176-77, 286, 310, 318, 347, 351. Recognizing the potential impact of low oil prices on Exxon Mobil’s classification of proved reserves, the 2015 10-K stated the following:

When crude oil and natural gas prices are in the range seen in late 2015 and early 2016 for an extended period of time, under the SEC definition of proved reserves, certain quantities of oil and natural gas, such as oil sands operations in Canada and natural gas operations in North America could temporarily not qualify as proved reserves. Amounts that could be required to be de-booked as proved reserves on an SEC basis are subject to being re-booked as proved reserves at some point in the future when price levels recover, costs decline, or operating efficiencies occur.

Under the terms of certain contractual arrangements or government royalty regimes, lower prices can also increase proved reserves attributable to ExxonMobil. We do not expect any temporary changes in reported proved reserves under SEC definitions to affect the operation of the underlying projects or to alter our outlook for future production volumes.

*Id.* ¶ 285. Despite the cautionary language in the 2015 10-K, Plaintiff asserts that the document was misleading, violated SEC disclosure standards, and was inconsistent with GAAP because the warning was too tepid considering prevailing oil prices and the high likelihood of a future de-booking. *Id.* ¶¶ 17, 180, 231, 348. Plaintiff further alleges

that, as 2016 progressed and oil prices continued to sag, the probability of an end-of-year de-booking became more certain, but Exxon Mobil's 2016 Form 10-Q reports—filed on May 4, 2016, and August 3, 2016—still failed to adequately inform investors of this reality. *Id.* ¶¶ 22, 169, 178-80, 184, 231, 348. Plaintiff alleges that these reports were misleading and violated SEC disclosure requirements. *Id.* ¶¶ 256, 348-52.

## **2.2. Exxon Mobil's RMDG Operations**

Plaintiff alleges that Defendants similarly misled the market about reserves tied to Exxon Mobil's RMDG Operations. Faced with declining domestic natural gas reserves, Exxon Mobil acquired XTO Energy, Inc., in December 2009. *Id.* ¶¶ 114, 116. The all-stock deal, valued between \$36 billion and \$41 billion, made Exxon Mobil the largest domestic natural gas producer in the United States. *Id.* ¶¶ 117, 120.

The deal proved inopportune for Exxon Mobil, as oil and gas prices began their extended decline in 2014. *Id.* ¶¶ 148-54. Between February 2014 and December 2015, the Henry Hub benchmark price—the most commonly used benchmark for natural gas produced in the United States—dropped 80%, from \$8.15 per million British thermal units (“BTU”) to \$1.63 per million BTU. *Id.* ¶¶ 59, 121, 155. Due in large part to that decline, Exxon Mobil's competitors in the natural gas sector, including those with dry gas operations in the Rocky Mountains, recorded impairment charges in 2014 and 2015. *E.g., id.* ¶¶ 156-57, 159, 161-64. Exxon Mobil did not follow suit. *Id.* ¶¶ 166-68. Plaintiff contends that Exxon Mobil's executives instead made misstatements to the

market in an attempt to present itself as unaffected by falling prices. *E.g., id.* ¶¶ 128-31, 134-37, 139, 192, 247, 270-71, 275-76.

Plaintiff asserts that by the end of 2015, the combination of low gas prices and prevailing market trends required Defendants to assess whether the RMDG reserves were impaired. *Id.* ¶¶ 169, 185-86. According to Plaintiff, had Defendants conducted a proper test, they would have recognized that the carrying value of the RMDG assets was no longer recoverable, necessitating an impairment charge. *Id.* ¶¶ 185, 191-93, 368-72. Because Exxon Mobil failed to properly recognize an impairment charge, Plaintiff contends, Exxon Mobil's 2015 10-K and certain subsequent 2016 Form 10-Q reports contained misleading financial information and violated SEC disclosure requirements. *Id.* ¶¶ 366-70, 373-76.

### **2.3. Proxy Costs of Carbon**

Plaintiff alleges that Defendants misled investors by not adhering to their public statements about using a proxy cost of carbon—a figure representing the projected effects of various climate-related policies on the future global energy demand—in their investment and valuation procedures. *E.g., id.* ¶¶ 7-8, 137-47. To reassure investors that Exxon Mobil would address climate change-related risks to its business, like increasing demand for renewable energy, in March 2014, Defendants issued a public report stating that Exxon Mobil would use a proxy cost of carbon. *Id.* ¶¶ 3, 295.

Plaintiff contends that Exxon Mobil used a lower proxy cost of carbon or no proxy cost of carbon in its internal proved reserves and impairment calculations for the



Canadian Bitumen Operations, including Kearl, and in asset impairment evaluations for the RMDG Operations. *E.g., id.* ¶¶ 137-47, 191-92, 243-45, 361, 371. According to Plaintiff, had Exxon Mobil properly employed its publicly declared proxy costs of carbon to assess the profitability of the Kearl and RMDG Operations, its eventual de-booking and impairment announcements would have been made far sooner. *See id.* ¶¶ 137-47, 175-77, 191-94. Plaintiff largely bases these allegations on the Affirmation of John Oleske (the “Oleske Affirmation”), which was filed on June 2, 2017, as part of an investigation by the New York Attorney General (“NYAG”) into whether Exxon Mobil misled the public and investors regarding the business risks associated with climate change. *E.g., id.* ¶¶ 137-47, 191-92, 243-45, 361, 371; *see* Doc. No. 36-1.

## **2.4. The Alleged Corrective Disclosures**

Plaintiff alleges that Defendants’ statements about Exxon Mobil’s Canadian Bitumen Operations, RMDG Operations, and Exxon Mobil’s use of a proxy cost of carbon were revealed to be misleading in a series of seven corrective disclosure (the “Corrective Disclosures”) dated November 9, 2015; January 20, 2016; July 29, 2016; August 10, 2016; October 28, 2016; January 18, 2017; and January 31, 2017. Doc. No. 36 at 167-72; Doc. No. 98-8 at 6.

### **2.4.1. The First Corrective Disclosure: The Guardian Article (November 9, 2015)**

On November 9, 2015, *The Guardian* reported that the NYAG was investigating Exxon Mobil for providing false information to the public regarding climate change and the potential business risks associated with it, including whether Exxon Mobil funded

“climate denial front groups” and “spread[] disinformation about climate science.”  
Doc. No. 36 ¶ 426; Doc. No. 98-5.

**2.4.2. The Second Corrective Disclosure: *The Los Angeles Times* Article (January 20, 2016)**

On January 20, 2016, *The Los Angeles Times* reported that the California Attorney General (“CAAG”) was “investigating whether Exxon repeatedly lied to the public and investors about the risks to its business from climate change, specifically whether Exxon’s actions ‘could amount to securities fraud and violations of environmental laws.’” Doc. No. 36 ¶ 429; Doc. No. 98-6.

**2.4.3. The Third Corrective Disclosure: Second Quarter Earnings Announcement for 2016 (July 29, 2016)**

On July 29, 2016, at 8:00 a.m. ET, Exxon Mobil released its second quarter earnings for 2016. Doc. No. 98-7. According to Plaintiff, Exxon Mobil revealed “a significant miss of expectations in Upstream”—the segment of Exxon Mobil’s business involving the exploration, acquisition, development, and extraction of unprocessed oil and gas commodities—“including a reported loss of \$514 million in U.S. Upstream . . . driven heavily by poor performance by XTO and Kearn.” Doc. No. 98-8 at 6; *see* Doc. No. 36 ¶ 40.

**2.4.4. The Fourth Corrective Disclosure: *The Washington Post* Op-Ed (August 10, 2016)**

On August 9, 2016, *The Washington Post* published an op-ed by Senators Elizabeth Warren and Sheldon Whitehouse entitled *Big Oil’s Master Class in Rigging the System*. Doc. No. 36 ¶ 432; Doc. No. 98-10 at 2-4. Plaintiff alleges the op-ed revealed

that “Exxon and its allies with financial ties to the oil and gas industry were harassing and bullying investigators in an attempt to ‘sidetrack state investigations and silence groups petitioning the government to address [Exxon’s] potential wrongdoing’ and avoid ‘court-supervised discovery . . . into whether it has spent decades deliberately deceiving the public about the harms associated with [climate change].” Doc. No. 36 ¶ 432.

Plaintiff also references a report from Environment and Energy Publishing LLC (the “EEP Report”), published on August 10, 2016. Doc. No. 36 ¶¶ 433-34. According to Plaintiff, the report detailed the calls of certain politicians for “Exxon executives to testify about climate change in light of state Attorney General investigations into whether Exxon knowingly misled the public and investors regarding the risks of carbon emissions.” *Id.*

#### **2.4.5. The Fifth Corrective Disclosure: Third Quarter Earnings (October 28, 2016)**

On October 28, 2016, at 8:00 a.m. ET, Exxon Mobil released its third quarter earnings for 2016. *Id.* ¶ 437; Doc. No. 88-3. Plaintiff characterizes the earnings release as disclosing that Exxon Mobil “might be forced to de-book nearly 20% of its oil and gas reserves, specifically acknowledging that it might have to de-book 3.6 billion barrels of oil sand reserves and one billion barrels of other North American reserves” if energy prices did not improve. Doc. No. 36 ¶ 437; Doc. No. 88-3 at 6.

#### **2.4.6. The Sixth Corrective Disclosure: UBS Downgrade (January 18, 2017)**

On January 18, 2017, after the close of trading, UBS downgraded Exxon Mobil to “sell” and reduced its price target from \$86 to \$77. Doc. No. 36 ¶ 443. Plaintiff notes that UBS cited Exxon Mobil’s “risk of de-booking up to 4.6 of its 24.8 BBoe of proved reserves” in its report downgrading the Company. *Id.*; Doc. No. 88-7.

#### **2.4.7. The Seventh Corrective Disclosure: Fourth Quarter Earnings Announcement for 2016 (January 31, 2017)**

On January 31, 2017, Exxon Mobil released its fourth quarter earnings for 2016. Doc. No. 36 ¶¶ 446-47. The Company confirmed that it would be taking an asset impairment charge of about \$2 billion largely related to the RMDG Operation and that it would de-book the Kearl Operation reserves within the coming weeks. *Id.*; Doc. No. 88-4.

### **3. LEGAL STANDARDS**

Before a court can certify a class, the party seeking class certification bears the burden of proving by a preponderance of the evidence that the class meets all four requirements of Fed. R. Civ. P. 23(a): (1) numerosity, (2) commonality, (3) typicality, and (4) adequacy. *Mary Kay Inc. v. Reibel*, 327 F.R.D. 127, 129 (N.D. Tex. 2018) (Fitzwater, J.) (citing *Alaska Elec. Pension Fund v. Flowserve Corp.*, 572 F.3d 221, 228 (5th Cir. 2009)). The court must conduct a “rigorous analysis” of each Rule 23(a) factor. *Wal-Mart Stores, Inc. v. Dukes*, 564 U.S. 338, 351 (2011). The party seeking class certification must establish by a preponderance of the evidence that the class

meets the requirements of at least one subsection of Rule 23(b). *Mary Kay Inc.*, 327 F.R.D. at 129 (citing *Flowserve*, 572 F.3d at 228). For proposed classes seeking monetary damages, like this one, the relevant requirements are “predominance” and “superiority” under Rule 23(b)(3). *Unger v. Amedisys Inc.*, 401 F.3d 316, 320 (5th Cir. 2005).

Although a court does not reach the merits of the case in evaluating whether class treatment is appropriate, it may look past the pleadings to understand the claims, defenses, relevant facts, and applicable substantive law in order to make a meaningful decision on class certification. *Castano v. Am. Tobacco Co.*, 84 F.3d 734, 744 (5th Cir. 1996).

## 4. ANALYSIS

### 4.1. Rule 23(a)

#### 4.1.1. Numerosity

To establish numerosity under Rule 23(a)(1), Plaintiff must show that its proposed “class is so numerous that joinder of all members is impracticable.” In examining numerosity, the Court considers factors such as “the geographical dispersion of the class, the ease with which class members may be identified, [and] the nature of the action.” *Zeidman v. J. Ray McDermott & Co.*, 651 F.2d 1030, 1038 (5th Cir. 1981) (quoting *Philips v. Joint Legis. Comm.*, 637 F.2d 1014, 1022 (5th Cir. 1981)). Defendants do not dispute that Plaintiff has satisfied Rule 23(a)(1).

The Court finds that Plaintiff has sufficiently established numerosity. Exxon Mobil's stock is a nationally traded security, with over four billion shares outstanding and an average weekly trading volume of tens of millions of shares. Doc. No. 88-1 ¶¶ 23-24, 29-30; *id.* at 108. The massive number of outstanding shares and the average weekly rate at which these shares are traded suggests that the purchasers of shares comprising Plaintiff's proposed class are too numerous to practicably join. *See Zeidman*, 651 F.2d at 1039 (collecting cases). The fact that Exxon Mobil trades nationally on the NYSE serves as additional evidence of numerosity, as it is highly likely that the individuals or entities involved in trading Exxon Mobil securities are spread throughout the country. *Id.*

#### **4.1.2. Commonality**

To establish commonality under Rule 23(a)(2), Plaintiff must demonstrate that "there are questions of law or fact common to the class." Even "a single common question will do." *Dukes*, 564 U.S. at 359. "The test for commonality is not demanding and is met 'where there is at least one issue, the resolution of which will affect all or a significant number of the putative class members.'" *Mullen v. Treasure Chest Casino, LLC*, 186 F.3d 620, 625 (5th Cir. 1999) (quoting *Lighbourn v. County of El Paso*, 118 F.3d 421, 426 (5th Cir. 1997)). Defendants do not dispute that Plaintiff has demonstrated commonality.

The Court concludes that the commonality requirement is satisfied. Common questions in this case include whether Defendants made material misrepresentations

or omissions, whether the Individual Defendants controlled the content and dissemination of the allegedly misleading misrepresentations, and whether the putative class members incurred damages. *See* Doc. No. 87 at 15. These issues, central to the securities fraud claims of the proposed class, arise from the same basic set of facts and can be addressed through evidence common to the proposed class.

#### 4.1.3. Typicality

To establish typicality under Rule 23(a)(3), Plaintiff must show that “the claims or defenses of the representative parties are typical of the claims or defenses of the class.” This undemanding test “focuses on the similarity between the named plaintiffs’ legal and remedial theories and the theories of those whom they purport to represent.” *Lightbourn*, 118 F.3d at 426. “Typicality does not require complete identity of claims, but requires that the representatives’ claims share the same essential characteristics with the class members’ claims. Factual differences do not defeat typicality if the claims arise from a similar course of conduct and share the same legal theories.” *In re Elec. Data Sys. Corp. Sec. Litig.*, 226 F.R.D. 559, 565 (E.D. Tex. 2005) (citing *James v. City of Dallas*, 254 F.3d 551, 570 (5th Cir. 2001)).

Defendants argue that Plaintiff’s claims are atypical of the proposed class claims because Plaintiff purchased all of its Exxon Mobil stock after the first two alleged Corrective Disclosures, subjecting it to the unique defense that it could not have relied on any supposed misrepresentations that were corrected before its purchase. Doc. No. 115 at 35.



Class certification will not be denied, however, on the basis of the presence of a unique defense unless there is a significant risk that Plaintiff will unduly devote resources to litigating the defense that should be devoted to litigating other matters material to the proposed class. *See Lehocky v. Tidel Techs., Inc.*, 220 F.R.D. 491, 502 (S.D. Tex. 2004) (Hittner, J.). Defendants have not sufficiently established that the timing of Plaintiff's purchases of Exxon Mobil stock—even if unique to Plaintiff—poses such a risk. Plaintiff's claims are founded on the same factual allegations and legal theories as the claims of all class members: that Defendants' misstatements and omissions artificially inflated the price of Exxon Mobil's stock, and that investors suffered damages when the truth was revealed. Regardless, Defendants' unique defense argument is moot based on the Court's narrowed class definition, discussed in Section 5 below. The Court finds that the typicality requirement has been met.

#### **4.1.4. Adequacy**

Rule 23(a)(4) requires Plaintiff to demonstrate that it “will fairly and adequately protect the interests of the class.” In determining adequacy, courts consider three primary questions: (1) whether there are any conflicts of interest between the representative parties and the class they seek to represent; (2) whether the representative parties have the willingness and ability to play an active role in the litigation; and (3) whether class counsel has the competence and zeal to represent the class and protect the interests of the absentee class members. *Feder v. Elec. Data Sys.*

*Corp.*, 429 F.3d 125, 130 (5th Cir. 2005); *Berger v. Compaq Comput. Corp. (Berger I)*, 257 F.3d 475, 479 (5th Cir. 2001).

Plaintiff's interests align with the interests of the proposed class because both have suffered losses as a result of the same allegedly unlawful conduct. Plaintiff is highly motivated to pursue this case and maximize the recovery not only for itself, but also for the absentee class members. The Court also finds that Robbins Geller is adequate to serve as class counsel given its extensive experience in litigating securities class actions in federal court, *see* Doc. No. 88-13 at 5-9, 26-46, 100, and its diligent representation of Plaintiff thus far. Defendants have not presented any evidence of a conflict of interest between Plaintiff and the proposed class, and they have not raised any concerns regarding the adequacy of Robbins Geller as class counsel.

Defendants question the adequacy of Plaintiff based on the deposition testimony of its designated representative, Mr. Michael Swiderski, however. Doc. No. 115 at 36-39. Defendants contend that certain portions of Mr. Swiderski's testimony indicate that he does not completely understand the allegations against Defendants, that he does not know the current status of this litigation, and that what he does know about the litigation comes solely from Robbins Geller. *Id.*; *see* Doc. No. 104 at 33-34, Doc. No. 105-8. Defendants conclude that Plaintiff impermissibly delegated this case to Robbins Geller and is, "at most, a disinterested spectator, lacking even basic knowledge about this case." Doc. No. 115 at 37.

The Private Securities Litigation Reform Act of 1995 (“PSLRA”), 15 U.S.C. § 78u-4, “raises the standard adequacy threshold” for securities class actions by requiring that they “be managed by active, able class representatives who are informed and can demonstrate they are directing the litigation.” *Berger I*, 257 F.3d at 483. The PSLRA does not create any additional Rule 23(a)(4) adequacy requirements. *Berger v. Compaq Comput. Corp.* (*Berger II*), 279 F.3d 313 (5th Cir. 2002). Class representatives are not required to have a complete knowledge of the case. *Lehocky* 220 F.R.D. at 503 (citing *Baffa v. Donaldson, Lufkin & Jenrette Sec. Corp.*, 222 F.3d 52, 62 (2d Cir. 2000)). They “need not be legal scholars and are entitled to rely on counsel.” *Berger I*, 257 F.3d at 483. Class representatives do “need to know more than that they were ‘involved in a bad business deal.’” *Id.* at 483 (quoting *Kelley v. Mid-America Racing Stables, Inc.*, 139 F.R.D. 405, 410 (W.D. Okla. 1990)).

Although Defendant’s adequacy concerns are not entirely unfounded, the Court concludes that Plaintiff is willing and able to adequately participate in and direct this litigation. In his deposition testimony, Mr. Swiderski could not comment on whether this lawsuit has anything to do with climate change, he was seemingly unfamiliar with some of the alleged Corrective Disclosures, and he could not name any Defendant except Exxon Mobil. Doc. No. 105-8 at 7, 9, 11. Mr. Swiderski’s testimony also revealed that he understands the basic theory of liability in this case and that the Fund allegedly suffered damages because of misstatements and omissions by Defendants. *Id.* at 3, 9. Mr. Swiderski also demonstrated that he knows more than that the Fund was

purportedly involved in a bad business deal. For example, he testified (1) that he had devoted at least sixty hours to fulfilling his responsibilities as a proposed class representative, a majority of which happened before he received a deposition notice; (2) that he regularly communicates with counsel about this case; (3) that at least Plaintiff's liaison counsel reviewed the Complaint before it was filed; and (4) that his responsibilities are to "vigorously fight the case for the common good of the class members." *Id.* at 4, 6, 12, 13; *see* Doc. No. 25-2 at 8-9.

#### **4.2. Rule 23(b)(3)**

Having determined that Plaintiff has satisfied the prerequisites of Rule 23(a), the Court now proceeds to evaluate whether it meets the predominance and superiority requirements of Rule 23(b)(3).

##### **4.2.1. Predominance**

The predominance element of Rule 23(b)(3) requires that "questions of law or fact common to class members predominate over any questions affecting only individual members." As previously discussed in Section 4.1.2 above, the proposed class claims present a number of common questions. The predominance requirement is "far more demanding" than commonality, as it tests "whether proposed classes are sufficiently cohesive to warrant adjudication by representation." *Unger*, 401 F.3d at 320 (quoting *Amchem Prods., Inc. v. Windsor*, 521 U.S. 591, 623-24 (1997)). In evaluating predominance, courts consider whether "members of a proposed class will need to present evidence that varies from member to member," or whether "the same

evidence will suffice for each member to make a prima facie showing [or] the issue is susceptible to generalized, class-wide proof.” *Tyson Foods, Inc. v. Bouaphakeo*, 577 U.S. 442, 453 (2016) (quoting 2 W. Rubenstein, *Newberg on Class Actions* § 4:50, pp. 196-197 (5th ed. 2012)).

The Court must consider predominance on a claim-by-claim basis. *Prantil v. Arkema Inc.*, 986 F.3d 570, 577 (5th Cir. 2021) (citing *Castano*, 84 F.3d at 744). The predominance analysis therefore begins with the elements of the underlying cause of action. *Erica P. John Fund, Inc. v. Halliburton Co. (Halliburton I)*, 563 U.S. 804, 809 (2011). Plaintiff alleges violations under §§ 10(b) and 20(a) of the Exchange Act and SEC Rule 10(b)-5. Doc. No. 36 at 176-79. The elements of a § 10(b) and SEC Rule 10(b)-5 claim are: “(1) a material misrepresentation or omission by the defendant; (2) scienter; (3) a connection between the misrepresentation or omission and the purchase or sale of a security; (4) reliance upon the misrepresentation or omission; (5) economic loss [“damages”]; and (6) loss causation.” *Amgen Inc. v. Conn. Ret. Plans and Tr. Funds*, 568 U.S. 455, 460-61 (2013) (quoting *Matrixx Initiatives, Inc. v. Siracusano*, 563 U.S. 27, 37-38 (2011)). Control person liability under § 20(a) requires the existence of an independent “primary” violation of the securities laws—in this case, § 10(b) and SEC Rule 10(b)-5. *See Jacobowitz v. Range Res. Corp.*, 596 F. Supp. 3d 659, 690 (N.D. Tex. 2022) (Pittman, J.) (citing *Emps.’ Ret. Sys. v. Whole Foods Mkt., Inc.*, 905 F.3d 892, 905 (5th Cir. 2018)); *In re Dynegy, Inc. Sec. Litig.*, 226 F.R.D. 263, 286 n.69 (S.D. Tex. 2005). The individual defendant must have had actual power over the controlled

person and induced them, directly or indirectly, to commit the acts constituting the primary violation. *In re BP P.L.C. Sec. Litig.*, 843 F. Supp. 2d 712, 791 (S.D. Tex. 2012) (Ellison, J.) (citing *Dennis v. General Imaging, Inc.*, 918 F.2d 496, 509 (5th Cir. 1990)).

In applying the Rule 23(b)(3) predominance test to Plaintiff's claims, the Court asks whether "*questions common to the class predominate*," not whether those "questions will be answered, on the merits, in favor of the class." *Amgen*, 568 U.S. at 459 (emphasis added).

As to most of the elements of Plaintiff's claims, there are no meaningful individualized questions that predominate over the common questions. First, whether Defendants' alleged misrepresentations and omissions would have influenced the decision-making process of a reasonable investor—*i.e.*, whether they are material—is a common question. *Id.* Even if Plaintiff is unable "to prove materiality [that] would not result in individual questions predominating," it would simply end the case for the entire class. *Id.* at 459-60. The same is true for the scienter, falsity, and—for purposes of § 20(a)—the control, influence, and culpability elements of Plaintiff's claims. All of these elements depend on the conduct of Defendants and do not vary with the identity of the putative class member asserting the claims. Finally, Plaintiff's proposed out-of-pocket damages methodology computes damages on a class-wide basis because the methodology provides a mechanical way of calculating each putative class member's damages based on the times at which they bought and sold Exxon Mobil shares. *E.g.*, Doc. No. 87 at 21, 27-28; *see Ludlow v. BP, P.L.C.*, 800 F.3d 674, 683 (5th Cir. 2015)

(Higginbotham, J.) (“In short, in order to certify a class, the damages methodology must be ‘sound’ and must ‘produce commonality of damages.’” (quoting *Comcast Corp. v. Behrend*, 569 U.S. 27, 37 (2013))).

Reliance is different. An individual plaintiff can, of course, prove reliance by “showing that he was aware of a company’s statement and engaged in a relevant transaction—*e.g.*, purchasing common stock—based on that specific misrepresentation.” *Halliburton I*, 563 U.S. at 810. In the context of a securities class action, like this one, with numerous potential class members, how can a plaintiff demonstrate each class member’s reliance? The Supreme Court addressed this issue in *Basic Inc. v. Levinson*, 485 U.S. 224 (1988).

#### **4.2.1.1. Fraud-on-the-Market Theory and the Presumption of Reliance**

Recognizing the unrealistic evidentiary burden that would result from requiring proof of direct reliance from each member of a proposed § 10(b) and SEC Rule 10(b)-5 class, the Supreme Court in *Basic* blessed the use of the “fraud-on-the-market” theory to create a rebuttable presumption of class-wide reliance. *Id.* at 241-47. The fraud-on-the-market theory posits that the price of a security traded in an efficient market reflects all publicly available information, including any material misrepresentations about the security. *Amgen*, 568 U.S. at 462. “Thus, courts may presume that investors trading in efficient markets indirectly rely on public, material misrepresentations through their ‘reliance on the integrity of the price set by the market.’” *Id.* (quoting *Basic*, 485 U.S. at 245).



To invoke *Basic*'s rebuttable presumption that class members relied on an alleged misrepresentation, a plaintiff must ultimately prove: "(1) that the alleged misrepresentation was publicly known; (2) that it was material; (3) that the stock traded in an efficient market; and (4) that the plaintiff traded the stock between the time the misrepresentation was made and when the truth was revealed." *Halliburton Co. v. Erica P. John Fund, Inc. (Halliburton II)*, 573 U.S. 258, 268 (2014). At the class certification stage, however, a plaintiff need only establish the first, third, and fourth elements, known as publicity, market efficiency, and market timing, respectively. *Goldman Sachs Grp., Inc. v. Ark. Tchr. Ret. Sys.*, 141 S. Ct. 1951, 1959 (2021). The Court leaves materiality "to the merits stage because it does not bear on Rule 23's predominance requirement." *Id.* (citing *Amgen*, 568 U.S. at 466-68).

There is no dispute that Defendants' alleged misrepresentations were publicly known. Defendants made them during earnings calls or in public filings and reports. *See* Doc. No. 98-12 at 18-19. As addressed in Section 5 below, the class definition here includes only persons who traded Exxon Mobil stock between the time of the alleged misrepresentations and the revelation that they were false, satisfying the marketing timing requirement.

That leaves the market efficiency requirement. In evaluating whether a stock traded in an efficient market, the Fifth Circuit considers a number of factors, including:

(1) the average weekly trading volume expressed as a percentage of total outstanding shares; (2) the number of securities analysts following and reporting on the stock; (3) the extent to which market makers and arbitrageurs trade in the stock; (4) the company's eligibility to file SEC registration Form S-3 (as opposed to Form S-1 or S-2); (5) the existence of empirical facts showing a cause and effect relationship between unexpected corporate events or financial releases and an immediate response in the stock price; (6) the company's market capitalization; (7) the bid-ask spread for stock sales; and (8) float, the stock's trading volume without counting insider-owned stock.

*Unger*, 401 F.3d at 323 (internal quotation marks omitted) (first citing *Cammer v. Bloom*, 711 F. Supp. 1264, 1286-87 (D.N.J. 1989); and then citing *Krogman v. Sterritt*, 202 F.R.D. 467, 477-78 (N.D. Tex. 2001) (Lynn, J.)). Significantly, "market efficiency is a matter of degree"; the "markets for some securities are more efficient than the markets for others." *Halliburton II*, 573 U.S. at 271-72.

The Court finds, and Defendants agree, that the market for Exxon Mobil stock is efficient. Doc. No. 176 at 48. Plaintiff has demonstrated, *inter alia*, that Exxon Mobil stock trades on the NYSE at a very high weekly trading volume and is widely covered by securities analysts. Doc. No. 87 at 22-23. Exxon Mobil's market capitalization is massive, at well over \$100 billion. *Id.* at 25. And the market efficiently absorbs Exxon Mobil-specific news and rapidly translates it into changes in stock price. *See id.* at 24; Doc. No. 88-1 at ¶¶ 47-50. As plaintiff's expert's market efficiency analysis demonstrates, the market for Exxon Mobil stock is among the most efficient for stocks listed on the NYSE and Nasdaq. *See, e.g.*, Doc. No. 88-1 ¶¶ 29-31, 35-37, 98-99, 101-03.

Plaintiff has therefore discharged its burden to show market efficiency, as well as publicity and market timing. *Basic*'s rebuttable presumption of class-wide reliance applies.

#### **4.2.1.1.1. Rebutting *Basic*'s Presumption of Reliance: Price Impact**

Defendants seek to rebut *Basic*'s presumption of class-wide reliance. They contend that their alleged misrepresentations had no impact on the price of Exxon Mobil stock. Because the “fundamental premise” behind *Basic*'s presumption of reliance is that, in an efficient market, “an investor presumptively relies on a misrepresentation *so long as it was reflected in the market price* at the time of his transaction,” Defendants' contention would cause “*Basic*'s fraud-on-the-market theory and presumption of reliance [to] collapse.” *Halliburton II*, 573 U.S. at 278 (emphasis added) (quoting *Halliburton I*, 563 U.S. at 812).

It is not always clear how a “price impact” argument like Defendants' will differ from a materiality argument—an inquiry reserved for the merits stage. *Goldman Sachs*, 141 S. Ct. at 1959 (citing *Amgen*, 568 U.S. at 466-68). If a misstatement is immaterial, one would not expect the exposure of its falsehood to have a negative impact on the price of the stock to which it relates. In practice, this may not matter much, as courts assessing price impact at the class certification stage “‘should be open to *all* probative evidence on that question—qualitative as well as quantitative—aided by a good dose of common sense.’” That is so regardless whether the evidence is also relevant to a merits

question like materiality.” *Id.* at 1961 (quoting *In re Allstate Corp. Sec. Litig.*, 966 F.3d 595, 613 n.6 (7th Cir. 2020)).

The defendant bears the burden of persuasion to demonstrate a lack of price impact by a preponderance of the evidence. *Id.* at 1963. If a defendant presents evidence to rebut *Basic*’s presumption of class-wide reliance, “[t]he district court’s task is simply to assess all the evidence of price impact—direct and indirect—and determine whether it is more likely than not that the alleged misrepresentations had a price impact.” *Id.* This is a familiar concept. In a typical civil action, for example, the plaintiff has the burden to prove its case by a preponderance of the evidence. The plaintiff will offer whatever it can to meet this burden, and the defendant, often through cross-examination and its own evidence, will attempt to chip away at the plaintiff’s case. Where the factfinder determines that the evidence is in “equipoise,” the plaintiff loses; it did not prove its case by a preponderance of the evidence. *See id.* Of course, the plaintiff’s job is easier where the defendant fails to defend. But the burden is still the plaintiff’s. Weak evidence—even coupled with the defendant’s silence—may not be enough to overcome it.

The Court therefore turns to the evidence of price impact, keeping in mind that Defendants bear the burden of persuading the Court, by a preponderance of the evidence, that there was no impact. *Id.* at 1963.

#### 4.2.1.1.2. The Parties' Event Studies

Plaintiff and Defendants both submit expert reports on the issue of price impact. In his report, Plaintiff's expert, Professor Frank C. Torchio ("Professor Torchio"), employs an "inflation-maintenance" theory of price impact. Under this theory, "a misrepresentation causes a stock price 'to *remain* inflated by preventing preexisting inflation from dissipating from the stock price.'" *Id.* at 1959 (quoting *FindWhat Inv. Grp. v. FindWhat.com*, 658 F.3d 1282, 1315 (11th Cir. 2011)). The resulting measure of damages is "the amount that the stock's price would have fallen 'without the false statement.'" *Id.* at 1961 (quoting *Glickenhau & Co. v. Household Int'l, Inc.*, 787 F.3d 408, 415 (7th Cir. 2015)). To estimate this amount, plaintiffs typically "point to a negative disclosure about a company and an associated drop in its stock price; allege that the disclosure corrected an earlier misrepresentation; and then claim that the price drop is equal to the amount of inflation maintained by the earlier misrepresentation." *Id.* (first citing *Glickenhau*, 787 F.3d at 413-417; and then citing *In re Vivendi, S.A. Sec. Litig.*, 838 F.3d 223, 233-37, 253-59 (2d Cir. 2016)). Professor Torchio takes this approach.

Defendants' expert, Dr. Allen Ferrell ("Dr. Ferrell"), likewise tests whether Defendants' alleged misrepresentations caused Exxon Mobil's stock to maintain an inflated price. Both Dr. Ferrell and Professor Torchio (collectively, the "Experts") conduct event studies to determine the extent to which the changes in Exxon Mobil's stock price may have resulted from the information revealed in the alleged Corrective

Disclosures as opposed to broader market and industry trends or random price fluctuations. *See* Doc. No. 88-1, Doc. No. 98-12, Doc. No. 103. The Experts agree that price fluctuations should be statistically significant before the Court rejects the possibility that the fluctuations are merely the result of random price movements. *See, e.g.*, Doc. No. 88-1 at 115-17, Doc. No. 98-12 ¶ 21. Neither the Experts nor the parties dispute the fitness of the regression models used by the Experts for their event studies. *See, e.g.*, Doc. No. 98-12 ¶ 22; Doc. No. 104 at 12.

Professor Torchio applied his model to analyze the reaction of Exxon Mobil's stock price to three of the alleged Corrective Disclosures: the earnings announcements dated July 29, 2016, October 28, 2016, and January 31, 2017. *See* Doc. No. 88-1 at ¶¶ 73-91; Doc. No. 103 ¶¶ 60-86, Doc. No. 104 at 12-13. Examining the cumulative stock price movement over a "two-day window" spanning two consecutive trading days following each alleged Corrective Disclosure, Professor Torchio found statistically significant negative price reactions associated with all three disclosures. *E.g.*, Doc. No. 88-1 ¶¶ 73-91, Doc. No. 103 ¶¶ 60-86.

Plaintiff does not present expert analysis of fluctuations in Exxon Mobil's stock price associated with the alleged Corrective Disclosures on November 9, 2015, January 20, 2016, August 10, 2016, or January 18, 2017, because, in Plaintiff's view, analyzing fluctuations in response to Exxon Mobil's earnings announcements is sufficient. Doc. No. 104 at 36. As support for this view, Plaintiff asserts that it is Defendants' "burden to prove no price impact." Doc. No. 176 at 197.

While Defendants bear the burden of persuasion to demonstrate a lack of price impact by a preponderance of the evidence, Plaintiff's conclusion does not follow from its premise. Defendants have analyzed fluctuations in Exxon Mobil's stock price associated with all of the Corrective Disclosures alleged by Plaintiff, and the Court must consider Defendants' analysis to comply with *Goldman Sachs'* directive that the Court be "open to *all* probative evidence" when assessing price impact. 141 S. Ct. at 1961. In *Erica P. John Fund, Inc. v. Halliburton Co. (Halliburton III)*—a case on which Plaintiff relies, *see* Doc. No. 104 at 11, 14, 16, 26, 30—Judge Lynn evaluated the parties' experts' price impact opinions with respect to many corrective disclosures that were not earnings announcements. 309 F.R.D. 251, 270-71, 273-74, 276-80 (N.D. Tex. 2015) (Lynn, J.).

Dr. Ferrell's analysis of Exxon Mobil's stock price revealed statistically significant negative price reactions to at least some of the seven alleged Corrective Disclosures. Examining a "close-to-open" event study window that encompasses stock price movements from the market close before an alleged Corrective Disclosure to the first market open after the alleged Corrective Disclosure, Dr. Ferrell found statistically significant negative price reactions only on July 29, 2016, and October 28, 2016. Doc. No. 98-12 ¶¶ 32-66. Examining a "close-to-close" event study window that encompasses stock price movements from the market close before an alleged Corrective Disclosure to the market close after the alleged Corrective Disclosure, Dr. Ferrell found



statistically significant negative price reactions on the same dates, as well as January 20, 2016, and January 18, 2017. *Id.*

Dr. Ferrell also analyzed the reaction of Exxon Mobil's stock price to Defendants' alleged misrepresentations. *Id.* ¶¶ 30-31. Using a close-to-close window, Dr. Ferrell found that no statistically significant positive reaction occurred on the date of any alleged misrepresentation. *Id.* Citing, *inter alia*, *IBEW Loc. 98 Pension Fund v. Best Buy Co.*, 818 F.3d 775, 782-83 (8th Cir. 2016), Defendants argue that the lack of a statistically significant stock price increase associated with the alleged misstatements demonstrates a lack of price impact sufficient to rebut *Basic*'s presumption of class-wide reliance. Doc. No. 115 at 17.

The inflation-maintenance theory, however, recognizes "that statements that merely maintain inflation already extant in a company's stock price, but do not add to that inflation, nonetheless affect a company's stock price." *In re Vivendi*, 838 F.3d at 256. Defendants do not meaningfully dispute that the information contained in the alleged misrepresentations analyzed by Dr. Ferrell coincided with the expectations of the market, making a statistically significant stock price increase unlikely. *See* Doc. No. 115 at 17, 19; Doc. No. 104 at 30. Further, the inflation-maintenance theory is widely accepted by the federal judiciary. Merritt B. Fox & Joshua Mitts, *Event-Driven Suits and the Rethinking of Securities Litigation* 77 (Eur. Corp. Governance Inst., Working Paper No. 656, 2022) (collecting cases); *see also Ludlow*, 800 F.3d at 674 (affirming

certification of class based in part on “stock price inflation” theory). The Court sees no persuasive reason to buck that trend here.

#### 4.2.1.1.3. Event Study Windows

Having determined that the Court must analyze whether Exxon Mobil’s stock price reacted to any of the alleged Corrective Disclosures with fluctuations indicating that Defendants’ alleged misstatements maintained the stock price at an inflated level, the Court turns to the subject of the Experts’ primary disagreement: the appropriate event window for assessing price impact.

Professor Torchio advocates for the use of a two-day window to measure price impact. *See, e.g.*, Doc. No. 103 ¶ 72. He notes that the proper event window duration depends on case-specific factors like the intricacy of the disclosure at issue, how widely the disclosure is shared among the investing public, how the company being studied responds to the market’s interpretations of the event being studied, and the information from analysts and commentators that surfaces after the event. Doc. No. 88-1 ¶ 152; Doc. No. 103 at 22-31 (collecting authorities). Professor Torchio argues that, in this case, studying the two-day period following each of the three alleged Corrective Disclosures he analyzed best captures the full effect of the news and commentary related to the alleged Corrective Disclosures. Doc. No. 88-1 ¶ 56; Doc. No. 103 ¶¶ 29-43, 78-86.

Dr. Ferrell submits that the proper event window depends on the timing of each alleged Corrective Disclosure. Doc. No. 98-12 ¶¶ 19-20; *see also id.* at 8-13 nn.9-26

(collecting authorities). According to Dr. Ferrell, using a close-to-open event window is best for analyzing the market reaction to alleged Corrective Disclosures made while the market was closed. *Id.* Dr. Ferrell opines that this targeted window best reflects the market's reaction to the disclosure without influence from possible confounding news or events that could occur during regular trading hours, particularly for a stock with a market as efficient as Exxon Mobil's. *Id.*; *see* Doc. No. 176 at 120-25. By contrast, Dr. Ferrell opines that using a close-to-close window is most appropriate for disclosures issued during trading hours. *E.g.*, Doc. No. 98-12 ¶ 19. Both Experts seemingly agree that the close-to-close window is the "standard" or "default" window. Doc. No. 176 at 74, 167-68; *see* Doc. No. 98-12 at 10 n.16.

Dr. Ferrell agrees that the use of a two-day window may be appropriate in cases involving, for example, an inefficient market, a corrective disclosure on an unknown date, or corrective disclosures made in close succession, but he disagrees that a two-day window is appropriate here. *E.g.*, Doc. No. 176 at 133-37. He opines that the market for Exxon Mobil stock is efficient, the dates of the Corrective Disclosures are known, and the Corrective Disclosures were not issued in close succession. *Id.* Dr. Ferrell reasons that any disclosure or analyst commentary repeating information previously shared in a corrective disclosure will have no additional price impact in the highly efficient market for Exxon Mobil stock because investors will have already traded on the information in the corrective disclosure. Doc. No. 98-12 ¶¶ 20, 58. Dr. Ferrell

concludes that a two-day event window is too long to appropriately measure price impact. *Id.* ¶ 20.

At least for class certification purposes, the Experts’ disagreement about whether a two-day window is appropriate in this case is only relevant with respect to the January 31, 2017, Corrective Disclosure. Using their preferred event windows, each Expert agrees that Exxon Mobil stock experienced a statistically significant price decline on July 29, 2016, and October 28, 2016—the only other dates Professor Torchio analyzed. *See* Doc. No. 98-12 ¶¶ 43, 55; Doc. No. 103 ¶¶ 62, 69-70.

The Supreme Court has yet to adopt “any particular theory of how quickly and completely publicly available information is reflected in the market price.” *Basic*, 485 U.S. at 248 n.28. Here, based on the information presented, the Court agrees with Dr. Ferrell’s analysis of the narrow issue disputed by the Experts: In this case, a two-day window is unsuitable for measuring price impact in an efficient market. *See Halliburton III*, 309 F.R.D. at 268-69. A two-day window could include extraneous market noise, making it more difficult to understand the effect of the alleged Corrective Disclosures on Exxon Mobil’s stock price. The Court sees no reason to doubt the general observation made by Dr. Ferrell that the efficient market here rapidly incorporates the disclosure of material information into the price of the stock prior to the publication of analyst commentary. Any commentary that does not itself present new information will not further impact the price of the stock. And, at least as to the alleged Corrective Disclosure on January 31, 2017, Plaintiff does not appear to contend that any new

information became available in analyst commentary published afterwards. *See* Doc. No. 103 ¶¶ 75-86; Doc. No. 104 at 23-25; Doc. No. 176 at 135-37.

#### 4.2.1.1.4. The Corrective Disclosures Analyzed

Defendants attempt to rebut the *Basic* presumption of class-wide reliance by demonstrating a lack of price impact associated with each of the seven alleged Corrective Disclosures. *See* Doc. No. 115. Defendants contend that the price impact inquiry does not necessarily end with the finding of a statistically significant negative price reaction. *Id.* at 21-27. According to Dr. Ferrell, “[a]n event study can tell us that something happened, but it can’t tell us *why*.” Doc. No. 98-12 ¶ 23 (quoting Bernard S. Black & Ronald J. Gilson, *THE LAW AND FINANCE OF CORPORATE ACQUISITIONS* 221 (Westbury NY: The Foundation Press, Inc., 1995)). To properly attribute a stock price movement to an alleged Corrective Disclosure, Defendants reason, the Court should consider, *inter alia*, the total mix of information the market possessed prior to the disclosure, the market reaction to similar, prior disclosures, and any additional confounding factors that might have impacted the stock price within the event window. *See, e.g.*, Doc. No. 115 at 17-28. Consistent with *Goldman Sachs*, the Court considers all evidence of price impact for each of the seven alleged Corrective Disclosures, “regardless whether that evidence overlaps with materiality or any other merits issue.” 141 S. Ct. at 1961.

4.2.1.1.4.1. November 9, 2015: *The Guardian* Article

The first alleged Corrective Disclosure came on November 9, 2015, at 10:30 a.m. ET, when *The Guardian* reported on the investigation by the NYAG into whether Exxon Mobil disseminated inaccurate information regarding climate change and its potential business ramifications. Doc. No. 36 ¶ 426; Doc. No. 98-5. Plaintiff did not specifically address this Corrective Disclosure in either its Motion or Reply, and Professor Torchio did not analyze or provide an opinion about it. *See* Doc. No. 87; Doc. No. 88-1; Doc. No. 103; Doc. No. 104; Doc. No. 176 at 94-95. Plaintiff ostensibly considers it partially corrective of some of Exxon Mobil's alleged misstatements about its use of a proxy cost of carbon.

The Court concludes that Defendants have rebutted the *Basic* presumption by a preponderance of the evidence by showing that the alleged November 9, 2015, Corrective Disclosure did not impact Exxon Mobil's stock price. Dr. Ferrell analyzed the reaction of Exxon Mobil's stock price to the November 9, 2015, article and found no statistically significant negative price reaction. Doc. No. 98-12 ¶¶ 35-38.

The Court finds further support for its conclusion in Dr. Ferrell's analysis of the reaction of Exxon Mobil's stock price to two additional articles. Dr. Ferrell analyzed the stock price reaction to a November 5, 2015, 6:38 p.m. ET, article published by *The Guardian* that is identical to the November 9, 2015 article. *Compare* Doc. No. 98-4, *with* Doc. No. 98-5. Dr. Ferrell also analyzed the stock price reaction to a November 5, 2015, 12:05 p.m. ET article published by the *New York Times* that conveyed the same

essential information as the articles published by *The Guardian*. Doc. No. 98-12 at 23-24 (citing Justin Gillis & Clifford Krauss, *Exxon Mobil Under Investigation in New York Over Climate Statements*, N.Y. TIMES (Nov. 5, 2015)).

Dr. Ferrell's analysis of Exxon Mobil's stock price reaction to the November 5, 2015, articles found no statistically significant price reaction on either a close-to-open or close-to-close basis. *Id.* ¶¶ 37-38. Dr. Ferrell's assessment of a sample of 480 contemporaneous research analyst reports revealed only six containing analyst commentary regarding the NYAG's investigation. *Id.* None of the six analyst reports regarded the NYAG's investigation as significant enough to necessitate adjustments to their price targets. *Id.*

#### **4.2.1.1.4.2. January 20, 2016: *The Los Angeles Times* Article**

The second alleged Corrective Disclosure is a report, published by *The Los Angeles Times* on January 20, 2016, at 3:00 a.m. ET, addressing an investigation by the CAAG into whether Exxon Mobil lied to the public and investors about the risks to its business from climate change. Doc. No. 36 ¶ 429; Doc. No. 98-6. Except to note that, on a close-to-close basis, Dr. Ferrell found that the alleged Corrective Disclosure had a statistically significant negative impact on Exxon Mobil's price, Plaintiff does not substantively address this alleged Corrective Disclosure in its Motion or Reply. *See* Doc. No. 87; Doc. No. 104 at 13, 28, 36. Professor Torchio also did not analyze or provide an opinion about the alleged Corrective Disclosure. *See* Doc. No. 88-1, Doc. No. 103; Doc. No. 104; Doc. No. 176 at 94-95. Nonetheless, Plaintiff deems the January 20,



2016, article to be partially corrective of some of Defendants' alleged misstatements about Exxon Mobil's use of a proxy cost of carbon. Doc. No. 104 at 28 n.19.

Given that the alleged January 20, 2016, alleged Corrective Disclosure occurred after the market was closed, Dr. Ferrell argues that any stock price reaction is best measured using a close-to-open window. *See* Doc. No. 98-12 ¶¶ 19-20, 40. Dr. Ferrell's close-to-open analysis revealed no statistically significant negative stock price reaction, though his close-to-close analysis did. *See id.* at 21.

If the use of a close-to-open window is ever appropriate, it would probably be for something like the alleged January 20, 2016, Corrective Disclosure—a short, easily digestible article issued hours before the market opened, discussing another investigation very similar to the one discussed in *The Guardian* and *The New York Times* articles mentioned above.

Considering the close-to-open analysis alongside other evidence, the Court is satisfied that Defendants have rebutted the *Basic* presumption by a preponderance of the evidence with respect to the January 20, 2016, Corrective Disclosure notwithstanding the statistically significant negative price reaction on a close-to-close basis. Dr. Ferrell's review of 480 concurrent research analyst reports found no analyst commentary regarding the CAAG's investigation. *Id.* ¶ 41. As with the NYAG investigation discussed in Section 4.2.1.1.4.1 above, the dearth of analyst commentary on the CAAG's investigation suggests that the market likely did not consider it to be significant. There is no indication that the market took the CAAG investigation more

seriously than the NYAG investigation, or that it saw political attention and multiple investigations as more troubling than one, as evidenced by the lack of a statistically significant negative price reaction associated with *The Washington Post* op-ed addressed below, which discussed another, similar investigation by the Attorney General of Massachusetts.

**4.2.1.1.4.3. July 29, 2016: Second Quarter Earnings Announcement for 2016**

On July 29, 2016, at 8:00 a.m. ET, Exxon Mobil issued the third alleged Corrective Disclosure, releasing second quarter earnings for 2016 wherein the company revealed an earnings miss in its upstream business. *See* Doc. No. 98-7. Even though the earnings release did not specifically attribute the miss to the Canadian Bitumen Operations (including Kearn proved reserves), the RMDG Operations, or carbon proxy costs, *see id.*, Plaintiff nevertheless alleges that it was partially corrective of misstatements related to all three. *See* Doc. No. 104 at 12-13, 15, 25-28.

The parties agree that Exxon Mobil stock experienced a statistically significant negative price movement on July 29, 2016, including on a close-to-open basis. *See* Doc. No. 98-12 ¶ 43; Doc. No. 104 at 12-13. Defendants argue that the movement is attributable to factors other than their alleged misrepresentations. *See* Doc. No. 98-12 ¶ 50. Alternatively, Defendants argue that, even if the earnings release was partially corrective of Defendants' alleged misrepresentations, the market did not realize that the release was corrective, negating any price impact stemming from the correction. *See* Doc. No. 115 at 23-26.

The Court agrees with Defendants and finds that they have rebutted the *Basic* presumption by a preponderance of the evidence by showing that the alleged July 29, 2016, Corrective Disclosure did not impact Exxon Mobil's stock price.

First, the Court accepts that the statistically significant negative price movement on July 29, 2016, was likely not attributable to the alleged corrective information in the earnings release about the Kearl Operation or the Canadian Bitumen Operations at large. This is because Imperial also released its quarterly earnings the morning of July 29, 2016 (at 7:55 a.m. ET), and, on a close-to-open basis, Imperial's stock price did not experience a statistically significant negative price reaction that day. *E.g.*, Doc. No. 98-12 ¶ 45. Recall that Imperial owns approximately 70% of Kearl and that, when Exxon Mobil issued its second quarter earnings release for 2016, the proved reserves at Kearl constituted a substantial part of the overall proved reserves from the Canadian Bitumen Operations, which in turn made up a significant portion of Exxon Mobil's total worldwide proved reserves. *E.g.*, Doc. No. 36 ¶¶ 96-101. If a downturn in performance at the Kearl Operation was responsible for Exxon Mobil's stock's statistically significant adverse close-to-open price movement on the day of the earnings release, one would therefore anticipate a comparably negative statistically significant close-to-open price movement in Imperial's stock. Because there was no such movement, the Court doubts that the observed reaction in Exxon Mobil's stock can reasonably be linked to Kearl and the broader Canadian Bitumen Operations.

As Dr. Ferrell explains, confounding factors likely explain the movement in Exxon Mobil's stock price on July 29, 2016. Reviewing twenty-four contemporaneous analyst reports, Dr. Ferrell found that analysts mainly ascribed the earnings shortfall to Canadian wildfires and civil disturbances in Nigeria, which collectively resulted in an estimated decrease in production of 100,000 barrels per day. Doc. No. 98-12 ¶ 50. Dr. Ferrell's review did not reveal any analyst commentary concerning asset impairments or possible de-bookings of Exxon Mobil's proved reserves. *Id.*

On cross-examination, Professor Torchio agreed with Dr. Ferrell "that there was no information that the market took from the earnings disclosure that connected to Kearl or Rocky Mountain." Doc. No. 176 at 104; *see also* Doc. No. 103 ¶ 112. Absent information in the earnings release about the Canadian Bitumen Operations (including Kearl) or the RMDG Operations, there should be no price impact from the release related to either Operation.

Finally, the Court notes that the evidence demonstrating that the July 29, 2016, earnings release did not, as relevant here, have an impact on Exxon Mobil's stock price is consistent with Plaintiff's theory of the case—that Defendants made misleading statements to present Exxon Mobil as unaffected by the market forces harming its competitors, and that they successfully kept the market in the dark until they finally revealed the truth to the market on October 28, 2016. *See* Doc. No. 104 at 18-19, 22.

**4.2.1.1.4.4. August 10, 2016: *The Washington Post* Op-Ed**

On the evening of August 9, 2016, *The Washington Post* published an op-ed by Senators Elizabeth Warren and Sheldon Whitehouse, entitled *Big Oil's Master Class in Rigging the System*. Doc. No. 36 ¶ 432; Doc. No. 98-10 at 2-3. This is the fourth alleged Corrective Disclosure. In the op-ed, Senators Warren and Whitehouse recapped the allegations that Exxon Mobil made misleading statements about climate change and the risks climate change posed to its long-term business model. Doc. No. 98-10 at 2. The Senators noted the existence of the NYAG investigation and a similar probe of Exxon Mobil being conducted by the Attorney General of Massachusetts. *Id.*

The Complaint also references the EEP Report, which was published on August 10, 2016, 1:11 p.m. ET. Doc. No. 36 ¶¶ 433-34. This report purportedly discussed certain politicians' desire for Exxon Mobil executives to testify in light of the multiple state attorneys general's investigations. *Id.*

Plaintiff does not specifically address *The Washington Post* op-ed or the EEP Report in its Motion or Reply, and Professor Torchio did not analyze the impact of either publication on Exxon Mobil's stock price. *See* Doc. No. 87; Doc. No. 88-1; Doc. No. 103; Doc. No. 104; Doc. No. 176 at 95. As with *The Guardian* and *The Los Angeles Times* articles, Plaintiff seemingly considers the publications to be partially corrective of some of Exxon Mobil's alleged misstatements about its use of a proxy cost of carbon.

The Court finds that Dr. Ferrell's analysis of the effect of the op-ed and the EEP Report on Exxon Mobil's stock price rebuts the *Basic* presumption of reliance. Because

the op-ed was published after trading hours on August 9, 2016, and the EEP Report was published during trading hours on August 10, 2016, Dr. Ferrell's analysis of the window from market close on August 9, 2016, to market close on August 10, 2016, accounted for both publications. Dr. Ferrell's analysis revealed no negative stock price reaction attributable the publications. *See* Doc. No. 98-12 at 21.

Not only does Dr. Ferrell's analysis sufficiently rebut the *Basic* presumption by showing a lack of price impact, but also it further supports Defendants' assertion that the market did not care about the state attorneys general's investigations into whether Exxon Mobil misled investors about its internal use of carbon proxy costs.

**4.2.1.1.4.5. October 28, 2016: Third Quarter Earnings Announcement for 2016**

On October 28, 2016, at 8:00 a.m. ET, Exxon Mobil issued the fifth alleged Corrective Disclosure, releasing its third quarter earnings for 2016. Doc. No. 36 ¶ 437; Doc. No. 88-3 at 6. Until this point, Plaintiff alleges, Defendants successfully misled investors into believing that Exxon Mobil was immune from the broader market forces impacting its peers. For example, Plaintiff cites an analyst report, issued five weeks prior to the alleged Corrective Disclosure, stating that Exxon Mobil's ability to steer clear of write-down issues was due to the company's unique conservative approach to the timing and extent of its capitalization of reserves. Doc. No. 104 at 18. According to Plaintiff, on October 28, 2016, Exxon Mobil "finally came clean about likelihood of reserve write-downs and impairments" when it revealed in its earnings announcement that it might have to de-book approximately 3.6 billion barrels of oil sand reserves and

1 billion barrels of other North American reserves. *Id.* at 19; *e.g.*, Doc. No. 36 ¶ 437; Doc. No. 88-3 at 6.

This is the Corrective Disclosure from which Defendants cannot escape. The parties agree that Exxon Mobil's stock experienced a statistically significant negative price reaction on October 28, 2016, including on a close-to-open basis. *See* Doc. No. 98-12 ¶ 55; Doc. No. 104 at 13.

Defendants primarily contend that the third quarter earnings release could not have had any relevant impact on Exxon Mobil's stock price because all meaningful information regarding Exxon Mobil's de-bookings, impairments, or losses at the Canadian Bitumen Operations was accessible to the market before the release. *E.g.*, Doc. No. 115 at 26-28. Defendants reason that the efficient market for Exxon Mobil's stock incorporated this information prior to the issuance of the release. *Id.*

There is reason to doubt Defendants' assertion that the market had access to the information about the Canadian Bitumen Operations disclosed in the earnings release before Exxon Mobil issued the release. Both parties' Experts cite analyst reports that express surprise and dismay about the potential de-bookings and impairments announced in the release. *See* Doc. No. 98-12 at 36-37.

If Defendants' theory were correct, one would expect to observe two results that did not, in fact, materialize. First, if the market could infer that Exxon Mobil's reserves were impaired or should be de-booked from competitors' impairments and de-bookings, one would expect to see a statistically significant negative price reaction for



Exxon Mobil's stock on the days when its peers announced their impairments and de-bookings. Professor Torchio's analysis demonstrates that the expected result did not occur. *See* Doc. No. 103 ¶¶ 140-43. Second, one would expect to observe no statistically significant negative movement in Exxon Mobil's stock price on October 28, 2016, unless the reaction was attributable to other factors. Since Exxon Mobil's stock price dropped on October 28, 2016, Defendants should be able to identify such factors. Dr. Ferrell suggests that the earnings release primarily emphasized the negative impact of lower refining margins and commodity prices on Exxon Mobil's results, while also spotlighting notable year-over-year drops in earnings and capital and exploration expenditures. Doc. No. 98-12 ¶ 57. As the contemporary analyst reports show, the market was concerned with de-booking and impairment; the Court finds that the factors identified by Dr. Ferrell do not fully explain the drop in Exxon Mobil's stock price on October 28, 2016. *See id.* at 36-37.

Even if Defendants are correct that all of the de-booking and impairment information disclosed in the third quarter earnings release was discoverable in advance of the issuance, Defendants have not necessarily shown that the information was so widely known that it had become an integral part of the total mix of information available to investors. *E.g., City of Roseville Emps. Ret. Sys. v. EnergySolutions, Inc.*, 814 F. Supp. 2d 395, 415 (S.D.N.Y. 2011). Defendants' theory that the market could have used publicly available information to deduce the information subsequently disclosed

in the earnings report largely relies on expert analysis. Doc. No. 115 at 26-28. As the Fifth Circuit has explained:

While it is generally true that in an efficient market, any information released to the public is presumed to be immediately digested and incorporated into the price of a security, it is plausible that complex economic data understandable only through expert analysis may not be readily digestible by the marketplace.

*Pub. Emps. Ret. Sys. of Miss. v. Amedisys, Inc.*, 769 F.3d 313, 323 (5th Cir. 2014).

The negative impact of the third quarter earnings release on Exxon Mobil's stock price does not, however, support Plaintiff's claim that Defendants misled the market about Exxon Mobil's use of a proxy cost of carbon. Plaintiff argues that the third quarter earnings release, together with the second and fourth quarter earnings releases also alleged to be Corrective Disclosures, is partially corrective of Defendants' purportedly misleading statements about carbon proxy costs. Doc. No. 104 at 9-10, 26-27. Plaintiff contends that the de-booking and impairments revealed on October 28, 2016, would have happened earlier had Exxon Mobil adhered to its publicly declared proxy cost. *Id.* Plaintiff alleges that if Exxon Mobil had incorporated its publicly stated carbon proxy cost in its internal models, it would have realized that its reserves were less profitable than it told the market. *See id.* According to Plaintiff, the three earnings releases are thus "linked" in a way that they can be seen as partially correcting the purported proxy cost misstatements. *Id.*

Plaintiff's "linkage" theory is unpersuasive. None of the three earnings releases alleged to be Corrective Disclosures pertained to Exxon Mobil's use of proxy costs of

carbon. With the exception of a single analyst report issued after the alleged October 28, 2016, Corrective Disclosure, which mentioned in passing the existence of the state attorneys general's investigations into Exxon Mobil, *see id.*, Plaintiff presents no evidence that the market connected the three earnings releases to Defendants' alleged misstatements about Exxon Mobil's use of a proxy cost of carbon, much less evidence that any connection between the releases and Defendants' alleged misstatements impacted the price of Exxon Mobil's stock.

The alleged Corrective Disclosures and similar articles that explicitly pertain to Exxon Mobil's use of carbon proxy costs (*i.e.*, *The New York Times* article, both of *The Guardian* articles, *The Los Angeles Times* article (on a close-to-open basis), *The Washington Post* article, and the EEP Report) had no statistically significant negative impact on Exxon Mobil's stock price. If any subsequent disclosure was corrective of Defendants' alleged carbon proxy cost misstatements, it would be the Oleske Affirmation, filed on June 2, 2017, at 7:50 a.m. ET, which unveiled all of the climate-related allegations pertinent to this case. Plaintiff does not classify the Oleske Affirmation as a corrective disclosure—perhaps because, as Dr. Ferrell demonstrates, the Affirmation did not produce a statistically significant negative change in Exxon Mobil's stock price on either a close-to-open or a close-to-close basis. *See* Doc. No. 98-12 at 21.

The Court concludes that Defendants have failed to rebut the *Basic* presumption by a preponderance of the evidence by showing that the alleged October 28, 2016, Corrective Disclosure had no impact on the price of Exxon Mobil stock, except insofar

as Plaintiff alleges that Defendants misled the market about Exxon Mobil's use of a proxy cost of carbon.

**4.2.1.1.4.6. January 18, 2017: UBS Downgrade**

On January 18, 2017, at 5:30 p.m. ET, UBS published the penultimate alleged Corrective Disclosure, an analyst report downgrading Exxon Mobil to "sell" and reducing the target price for its stock from \$86 to \$77. Doc. No. 36 ¶ 443; Doc. No. 88-7. According to Plaintiff, in downgrading Exxon Mobil and lowering the price target, UBS cited the likely risk that Exxon Mobil would de-book proved reserves at Kearl based on Exxon Mobil's third quarter earnings release. Doc. No. 87 at 12 n.6; Doc. No. 104 at 13, 28. Professor Torchio did not analyze or provide a price impact opinion about the UBS report. *See* Doc. No. 88-1, Doc. No. 103; Doc. No. 104; Doc. No. 176 at 95. Plaintiff, however, claims that the report is partially corrective of certain alleged misrepresentations in the 2015 10-K and other subsequent releases. *See* Doc. No. 87 at 12.

The Court concludes that Defendants have rebutted the *Basic* presumption by a preponderance of the evidence by showing a lack of price impact with respect to the UBS report. Dr. Ferrell's close-to-open analysis revealed no statistically significant negative stock price reaction, though his close-to-close analysis did. Doc. No. 98-12 at 21, 38. Because UBS issued the report after trading hours, Dr. Ferrell contends that a close-to-open window is the proper event window for assessing price impact. *Id.* ¶¶ 19-20, 61. Based on the information before the it, the Court finds that a close-to-open

window is probably well-suited to analyzing the impact of the UBS report on Exxon Mobil's stock price, as it was for analyzing the impact of *The Los Angeles Times* article on Exxon Mobil's stock price. UBS issued its brief report many hours before the market opened, and the market for Exxon Mobil stock is highly efficient.

The Court need not decide between close-to-open or close-to-close windows here, as the UBS report merely summarizes parts of Exxon Mobil's third quarter earnings release and makes a recommendation based on that information; it does not offer any new corrective information to the market. Plaintiff does not contend otherwise. *See* Doc. No. 104 at 28-29; Doc. No. 176 at 197 (“[T]he UBS downgrade is a result of Kearn de-booking and the lack of profitability.”). As previously mentioned, the Court agrees with Dr. Ferrell that the mere reiteration of already disclosed information should not influence the price of a stock traded in an efficient market.

**4.2.1.1.4.7. January 31, 2017: Fourth Quarter Earnings Announcement for 2016**

The final alleged Corrective Disclosure occurred on January 31, 2017, at 8:00 a.m. ET, when Exxon Mobil released its fourth quarter earnings for 2016 and confirmed that it would be taking an asset impairment charge of about \$2 billion largely related to the RMDG Operations. Doc. No. 36 ¶¶ 446-47; Doc. No. 88-4. Later that morning, in an earnings conference call that took place during the first ninety minutes of trading, Woodbury confirmed that the Kearn proved reserves discussed in the third quarter earnings announcement would be de-booked in the coming weeks. Doc. No. 36 ¶ 234; Doc. No. 104 at 23.

The Court concludes that Defendants have rebutted the *Basic* presumption by a preponderance of the evidence by showing a lack of price impact with respect to the fourth quarter earnings release. The fourth quarter release is the only earnings release alleged to be a Corrective Disclosure that is not associated with a statistically significant negative price reaction on a close-to-open basis or even a close-to-close basis. Dr. Ferrell analyzed Exxon Mobil's price using a close-to-open window and found no statistically significant negative price reaction to the fourth quarter release. Doc. No. 98-12 ¶¶ 63-64. Dr. Ferrell's close-to-close analysis—which would have captured any reaction associated with the de-booking confirmation in the conference call conducted after the market opened—also revealed no statistically significant negative price reaction. *See id.* Professor Torchio, did, however, find a statistically significant negative price reaction by using a two-day window. *E.g.*, Doc. No. 88-1 ¶¶ 83-91; Doc. No. 103 ¶¶ 75-86.

As the Court has already explained in Section 4.2.1.1.3 above, in this case, a two-day window is not appropriate for assessing price impact in the highly efficient market for Exxon Mobil's stock. The failure of the fourth quarter release to significantly affect Exxon Mobil's stock price when measured over a more appropriate window is unsurprising since the market was on notice of the impairments and de-bookings announced in the fourth quarter release as a result of Exxon Mobil's disclosures in its third quarter earnings release.

#### 4.2.1.2. Omissions-Based Presumption of Reliance

Apart from invoking the presumption of reliance stemming from the fraud-on-the-market theory, Plaintiff also claims that it is entitled to rely on the presumption of reliance first established in *Affiliated Ute Citizens of Utah v. United States*, 406 U.S. 128 (1972), to meet Rule 23(b)(3)'s predominance requirement. In *Affiliated Ute*, the Supreme Court held that proof of reliance is not a necessary condition for recovery in fraud cases primarily involving the omission of material information. *Id.* at 152-53. Following this precedent, the Fifth Circuit holds that in cases where the “defendant has failed to disclose any information whatsoever relating to material facts about which the defendant has a duty to the plaintiff to disclose,” a plaintiff is entitled to a presumption that the plaintiff relied on the omission. *Abell v. Potomac Ins. Co.*, 858 F.2d 1104, 1119 (5th Cir. 1988), *vacated on other grounds sub nom. Fryar v. Abell*, 492 U.S. 914 (1989); *see Regents of Univ. of Cal. v. Credit Suisse First Bos. (USA), Inc.*, 482 F.3d 372, 384-85 (5th Cir. 2007). The presumption does not apply “to cases where the plaintiffs allege either that the defendant has made false statements or has distorted the truth by making true but misleading incomplete statements.” *Steiner v. Southmark Corp.*, 734 F. Supp. 269, 276 (N.D. Tex. 1990) (Fitzwater, J.).

Plaintiff's claims involve a mix of allegedly fraudulent conduct, most of which appears to consist of misrepresentations rather than omissions. That mix may not support application of the *Affiliated Ute* presumption. *See Lehocky*, 220 F.R.D. at 510 (holding that plaintiffs could not rely on *Affiliated Ute* presumption where plaintiffs



alleged a mix of misrepresentations and omissions). But the Court need not and does not decide the applicability of *Affiliated Ute* to this case, as it has already found that Plaintiff is entitled to a rebuttable presumption of reliance based on *Basic*'s fraud-on-the-market theory.

#### 4.2.2. Superiority

Having found that there are questions of law and fact common to the class that predominate with respect to Plaintiff's non-carbon proxy cost allegations, the Court now turns to the question of whether the class action is the superior way of resolving this matter. Defendants do not contest the superiority of class treatment here. *See generally* Doc. No. 115.

Nonetheless, the Court independently examines whether a class action "is superior to other methods for fairly and efficiently adjudicating th[is] controversy" based on the factors in Rule 23(b)(3)(A)-(D):

- (A) the class members' interests in individually controlling the prosecution or defense of separate actions;
- (B) the extent and nature of any litigation concerning the controversy already begun by or against class members;
- (C) the desirability or undesirability of concentrating the litigation of the claims in the particular forum; and
- (D) the likely difficulties in managing a class action.

The motivation of any potential class members to individually prosecute separate actions appears minimal. Plaintiff was the sole entity to move for appointment as lead plaintiff in this matter. Doc. No. 87 at 29. There is no indication of any related

individual lawsuits; Plaintiff is not aware of any, and Defendants do not point to any either. *See id.*; Doc. No. 115. Considering the probable presence of numerous absent potential class members, each with relatively modest losses that might not justify the associated litigation costs and complexities, the desire for personal control over such cases seems relatively limited. *Id.*

In terms of efficiency, the potential class encompasses thousands of investors who purchased Exxon Mobil stock. The central claim for all investors is identical: Defendants misled them into acquiring Exxon Mobil stock at artificially inflated prices. As the elements of the putative class claims are issues that can be addressed uniformly for the proposed class, the Court believes that trying the claims collectively will be more efficient and cost-effective compared to individual trials.

Finally, the Court finds that the Northern District of Texas is a desirable forum for this class action at least because of the Court's familiarity with the case, and because a high number of the acts complained of occurred in substantial part in this District. *E.g., id.*

## **5. CLASS DEFINITION AND CONCLUSION**

The Court will certify Plaintiff's proposed class with modifications. "District courts have significant leeway and discretion over the management of class actions. They may modify the classes to fit the requirements better and should not dismiss an action purely because the proposed class definition is too broad." *Braidwood Mgmt., Inc. v. Equal Emp. Opportunity Comm'n*, 70 F.4th 914, 933-34 (5th Cir. 2023) (first citing

*Allison v. Citgo Petroleum Corp.*, 151 F.3d 402, 408 (5th Cir. 1998); and then citing *In re Monumental Life Ins. Co.*, 343 F.3d 331, 338 (5th Cir. 2003)).

As previously explained in Section 4.2.1.1.1, misrepresentations that do not impact the price of a stock are not actionable and cannot establish the starting point for a class period. The Court has determined that Defendants have sufficiently rebutted the alleged price impact of Defendants' alleged misstatements about the carbon proxy costs. The Court therefore **DENIES** Plaintiff's Motion to the extent it seeks to certify claims regarding Defendants' alleged misstatements about carbon proxy costs.

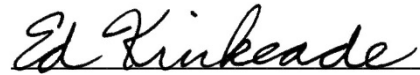
As outlined below, the Court certifies a class with respect to Plaintiff's remaining claims after adjusting the class period to reflect the Court's analysis of price impact. The Court sets the starting date of the class period on the date of the first alleged misstatement relevant to Plaintiff's certified claims: Exxon Mobil's 2015 10-K. As the end date for the class period, the Court sets October 28, 2016—the date of the sole alleged Corrective Disclosure whose impact on Exxon Mobil's stock price Defendants failed to rebut. Accordingly, the Court accordingly **CERTIFIES** the following class:

All persons who purchased or otherwise acquired Exxon Mobil Corporation common stock between February 24, 2016, and October 28, 2016 (the date of the sole alleged Corrective Disclosure whose impact on Exxon Mobil's stock price Defendants failed to rebut), inclusive, and were damaged thereby. Excluded from the class are Defendants and their families, the officers and directors of Exxon Mobil, at all relevant times, members of their immediate families and their legal representatives, heirs, successors or assigns, and any entity in which Defendants have or had a controlling interest.

The Court appoints Lead Plaintiff Greater Pennsylvania Carpenters Pension Fund as class representative and Robbins Geller Rudman & Dowd LLP as class counsel.

**SO ORDERED.**

Signed August 21<sup>st</sup>, 2023.

  
ED KINKEADE  
UNITED STATES DISTRICT JUDGE

# Exhibit 70

**UNITED STATES DISTRICT COURT  
MIDDLE DISTRICT OF FLORIDA  
TAMPA DIVISION**

JACOB J. BECKEL,

Plaintiff,

vs.

FAGRON HOLDINGS USA, LLC, JACOB G.  
JACKSON, FAGRON NV, GER VAN JEVEREN,  
and JAN PEETERS,

Defendants.

Case No. 8:16-cv-02059

**EXPERT REPORT**

**OF**

**LUCY P. ALLEN**

**November 16, 2018**

*CONFIDENTIAL*

## **I. SCOPE OF ASSIGNMENT**

1. I have been asked by counsel for Jacob J. Beckel to estimate damages to Mr. Beckel resulting from his receipt of shares of Fagron NV (“Fagron”) in connection with the sale of AnazaoHealth Corporation (“AnazaoHealth”). Mr. Beckel claims damages as a result of misrepresentations and omissions made by Defendants during the negotiations to acquire AnazaoHealth regarding Fagron’s involvement with billing practices for compound pain creams. For the purpose of this analysis, I am assuming liability as alleged in the Amended Complaint, filed October 31, 2018 (“Complaint”).

## **II. QUALIFICATIONS AND REMUNERATION**

### **A. Qualifications**

2. I am a Managing Director of NERA Economic Consulting (“NERA”) and a member of NERA’s Securities and Finance Practice. NERA provides practical economic advice related to highly complex business and legal issues arising from competition, regulation, public policy, strategy, finance, and litigation. NERA was established in 1961 and now employs approximately 500 people in more than 20 offices worldwide. NERA’s Securities and Finance Practice, which performs research in securities and financial markets, dates from the early 1970s and employs a research staff of more than 100 professionals holding degrees in economics, finance, and mathematics. The practice group counts among its clients major securities exchanges, risk managers, principals needing valuation services, and parties in litigation.

3. I have an A.B. from Stanford University, an M.B.A. with a concentration in Finance and Accounting from Yale University, and M.A. and M. Phil. degrees in Economics, also from Yale University. Prior to joining NERA, I was an Economist for both President George H. W. Bush’s and President Bill Clinton’s Council of Economic Advisers, providing economic analysis on regulation and health care policy issues. In my over 20 years at NERA, I have been engaged as an economic consultant or expert witness in numerous projects involving securities and financial economics. In the course of this work, I have analyzed the effect of information on stock prices of over 100 companies. My resume with recent publications and testifying experience is included as Appendix A.



## **B. Remuneration**

4. NERA is being compensated for time spent by me and my team at standard billing rates and for out-of-pocket expenses at cost. NERA currently bills for my time at \$900 per hour. NERA's fees are not in any way contingent upon the outcome of this matter.

## **III. MATERIALS CONSIDERED**

5. In preparing this report, I considered the following materials:

- a) Complaint;
- b) Plaintiff Jacob J. Beckel's Supplemental Answers to Fagron Defendants' First Set of Interrogatories, dated October 26, 2018;
- c) Plaintiff Jacob J. Beckel's Supplemental Answers to Jacob G. Jackson's First Set of Interrogatories, dated October 26, 2018;
- d) "Beckel SPA.pdf;"
- e) "Beckel Family ARSUF History.pdf;"
- f) "FAGRON\_00018226-00018266.pdf;"
- g) "FAGRON\_00018267-00018305.pdf;"
- h) "Draft Profile - AnazaoHealth 1-8-2014).pdf;"
- i) "Copy of AHC Budget-Forecast Summary-1.xlsx;"
- j) "Copy of AnazaoHealth Product Line Template v1 10.08.15-3.xlsx;"
- k) "AHC Budget-Forecast Summary;"
- l) Analyst reports on Fagron from Thomson Reuters;
- m) Fagron's annual reports, press releases, conference call transcripts and earnings presentations from the Fagron website (<https://investors.fagron.com/>);
- n) News stories on Fagron, AnazaoHealth, and the pharmaceutical industry from Factiva and Bloomberg, L.P.;

- o) Fagron's stock and trading volume data from Bloomberg, L.P.;
- p) Price data for market and industry indices from Bloomberg, L.P.;
- q) Price data and market capitalization for companies classified under NAICS code 446110 (Pharmacies and Drug Stores), SIC code 5912 (Drug Stores and Proprietary Stores) and FactSet industry code 3510 (Drugstore Chains) from FactSet Research Systems, Inc. ("FactSet");
- r) Yield on 3-month US Treasury Bills from the Federal Reserve Bank of St. Louis (<https://fred.stlouisfed.org/series/TB3MS>, accessed November 15, 2018); and
- s) Academic literature and textbooks on finance, securities, valuation and statistics.

## IV. BACKGROUND

### A. Company Background

6. Fagron is a pharmaceutical research and development company, based in Waregem, Belgium, focused on pharmaceutical compounding and personalized pharmaceutical care.<sup>1</sup> On May 11, 2015, Fagron announced that it had agreed to acquire AnazaoHealth, a compounding pharmacy specializing in nuclear, pain and intrathecal compounding (the "Acquisition").<sup>2</sup> Plaintiff Jacob J. Beckel was the Chief Executive Officer of AnazaoHealth at the time of the Acquisition.<sup>3</sup> According to the Complaint, Fagron and AnazaoHealth agreed on a purchase price of \$50 million, which included \$30 million in cash and \$10 million in shares of Fagron stock, and an additional \$10 million in Fagron shares as an earnout payment if AnazaoHealth met certain financial performance criteria following the Acquisition.<sup>4</sup>

---

<sup>1</sup> Fagron FY2015 Annual Report, p. 83. See, also, "Deal Enhances Sterile Compounding in the US," *Jefferies*, May 11, 2015.

Fagron stock is publicly traded on the Euronext Brussels and the Euronext Amsterdam stock exchanges. See, for example, Fagron FY2016 Annual Report, p. 4.

<sup>2</sup> "Press Release: Fagron NV: Acquisition of AnazaoHealth in the United States," *Dow Jones Institutional News*, May 11, 2015. See, also, Fagron FY2015 Annual Report, p. 127.

<sup>3</sup> Complaint, ¶12.

<sup>4</sup> "Press Release: Fagron NV: Acquisition of AnazaoHealth in the United States," *Dow Jones Institutional News*, May 11, 2015 and Complaint, ¶24.

7. At the time of the Acquisition, Fagron's business was divided into four segments – Fagron Specialty Pharma Services (“Services”), Fagron Essentials (“Essentials”), Fagron Trademarks (“Trademarks”), and HL Technology.<sup>5</sup> The Services segment, which provided sterile and non-sterile customized medication, and the Essentials segment, which supplied pharmaceutical raw materials, together accounted for the majority of Fagron's revenue and profits in FY2014 and FY2015.<sup>6</sup> In particular, these two segments accounted for over 84% of Fagron's revenue and REBITDA (*i.e.*, EBITDA excluding the impact of non-recurring items), a measure of Fagron's profitability, in both FY2014 and FY2015, as shown in the table below:

<b>Fagron FY2014 and FY2015 Revenue and REBITDA</b>					
	<b>FY2015</b>		<b>FY2014</b>		<b>Y-o-Y Growth</b>
	<b>Amount</b>	<b>% of Total</b>	<b>Amount</b>	<b>% of Total</b>	
<b>Total Revenue</b>	<b>€473.0M</b>	<b>100.0%</b>	<b>€447.1M</b>	<b>100.0%</b>	<b>+5.8%</b>
Services	€187.9M	39.7%	€147.8M	33.1%	+27.1%
Essentials	€225.2M	47.6%	€245.0M	54.8%	-8.1%
Trademarks	€50.3M	10.6%	€45.7M	10.2%	+10.3%
HL Technology	€9.5M	2.0%	€8.6M	1.9%	+11.3%
<b>Total REBITDA</b>	<b>€106.5M</b>	<b>100.0%</b>	<b>€118.5M</b>	<b>100.0%</b>	<b>-10.0%</b>
Services	€41.1M	38.6%	€43.3M	36.6%	-5.2%
Essentials	€48.6M	45.6%	€60.0M	50.6%	-19.0%
Trademarks	€15.6M	14.7%	€14.4M	12.2%	+8.5%
HL Technology	€1.2M	1.1%	€0.7M	0.6%	+70.2%
<b>Source:</b>					
Fagron FY2015 Annual Report.					

<sup>5</sup> Note that Fagron changed the name of the Services segment during FY2015 from Fagron Compounding Services to Fagron Specialty Pharma Services.

<sup>6</sup> Segment descriptions from Fagron FY2015 Annual Report, pp. 6-7.

## B. Summary of Allegations

8. Mr. Beckel alleges that Defendants made certain misrepresentations and omissions regarding Fagron's involvement with billing practices for compound pain creams during the negotiations to acquire AnazaoHealth.<sup>7</sup> For the purpose of this analysis, I have assumed Mr. Beckel's allegations are true. Mr. Beckel alleges the following:

- a) At the time of the negotiations, Mr. Beckel was aware that certain pharmacies that sold compound pain creams routinely overcharged health insurance companies (including TRICARE, the health insurance program for the US military) for these creams,<sup>8</sup> and those in the pharmaceutical compounding industry understood that this practice was "fraudulent" and "unsustainable,"<sup>9</sup> as insurance companies were "likely to reduce or eliminate reimbursements" to suppliers of compound pain creams due to overcharges.<sup>10</sup>
- b) During the negotiations, Mr. Beckel sought "information and assurances" from Defendants about "whether Defendants conducted any business relying on compound pain creams" and, if so, "whether Defendants engaged in or relied upon the inflated reimbursement/billing practice that the industry knew would collapse."<sup>11</sup>
- c) In response to Mr. Beckel's inquiries, Defendants "specifically and emphatically represented" that there was "no issue" with the manner in which Fagron or any related entity billed health insurers for compound pain creams.<sup>12</sup>
- d) Defendants further stated that Fagron charged \$200 to \$300 for compound pain creams, and that this pricing would be a "competitive advantage" and an "opportunity for increased profits" once insurance companies eliminated reimbursements to the

---

<sup>7</sup> Complaint, ¶¶60,69.

<sup>8</sup> Complaint, ¶18.

<sup>9</sup> Plaintiff Jacob J. Beckel's Supplemental Answers to Jacob G. Jackson's First Set of Interrogatories, dated October 26, 2018, pp. 2-3.

<sup>10</sup> Complaint, ¶18.

<sup>11</sup> Plaintiff Jacob J. Beckel's Supplemental Answers to Jacob G. Jackson's First Set of Interrogatories, dated October 26, 2018, p. 3.

<sup>12</sup> Complaint, ¶19 and Plaintiff Jacob J. Beckel's Supplemental Answers to Jacob G. Jackson's First Set of Interrogatories, dated October 26, 2018, p. 3.

- suppliers of compound pain creams that were engaged in “fraudulent” billing practices.<sup>13</sup>
- e) Defendants knew that their representations were false when made because, at the time of the negotiations, (i) Defendants were engaged in “fraudulent” billing practices and were charging health insurers “thousands of dollars” per compound pain cream, and (ii) Defendants were aware of the negative impact that changes in insurers’ reimbursement policies would have on Fagron’s business.<sup>14</sup>
- f) Mr. Beckel would not have agreed to sell AnazaoHealth in exchange for shares of Fagron stock if he had been aware of the “truth” regarding Fagron’s involvement with billing practices for compound pain creams.<sup>15</sup>

## V. METHODOLOGY

9. I estimated damages using two methods: (i) based on the amount by which the value of Fagron shares received by Mr. Beckel was inflated by the misrepresentations and omissions, and (ii) based on how the value of AnazaoHealth would have changed if it had not been acquired and had instead remained a standalone company.

10. For the first method, I estimated damages as the difference in value between what Mr. Beckel believed he was receiving in Fagron shares, and what he actually received in the Acquisition because of Defendants’ misrepresentations and omissions. This method entailed estimating the amount by which the value of the Fagron shares received by Mr. Beckel was inflated by the alleged misrepresentations and omissions. In other words, I estimated what percent of Fagron’s stock price at the time of the Acquisition was associated with the information regarding the company’s involvement with billing practices for compound pain creams that Defendants failed to disclose. To estimate the amount of inflation in the stock, I analyzed how the market and the stock price reacted when the information that Defendants failed

---

<sup>13</sup> Complaint, ¶19 and Plaintiff Jacob J. Beckel’s Supplemental Answers to Jacob G. Jackson’s First Set of Interrogatories, dated October 26, 2018, p. 3.

<sup>14</sup> Complaint, ¶21 and Plaintiff Jacob J. Beckel’s Supplemental Answers to Jacob G. Jackson’s First Set of Interrogatories, dated October 26, 2018, p. 4.

<sup>15</sup> Complaint, ¶¶34,53,62.

to disclose about Fagron's involvement with billing practices regarding compound pain creams, and about the negative impact of the changes in insurers' reimbursement policies on Fagron's business, was publicly announced.

11. For the second method, I estimated damages as the difference between what Mr. Beckel would have had today had he been told the information about Fagron's involvement with billing practices that Defendants failed to disclose and had not sold AnazaoHealth to Defendants, and what he actually received. This method entailed estimating how the value of AnazaoHealth would have changed between the time of the Acquisition and today if it had not been acquired and had instead remained a standalone company. I estimated the difference between what Mr. Beckel has today in Fagron stock and what he would have had if AnazaoHealth's value moved with the industry.

## **VI. ANALYSIS OF ALLEGED DAMAGES**

12. I estimated damages for the \$10.11 million worth of Fagron shares (equal to 226,700 shares valued at \$44.62 per share) that Mr. Beckel received in the Acquisition.<sup>16</sup> I find that the range of alleged damages is between \$2.51 million and \$4.15 million, as summarized in the table below:

---

<sup>16</sup> Number of Fagron shares received by Mr. Beckel from "Beckel Family ARSUF History.pdf." Share price from "Press Release: Fagron issues new shares to finance acquisition of AnazaoHealth," *Dow Jones Institutional News*, June 30, 2015.

It is my understanding that Mr. Beckel also received 50,000 warrants to acquire shares of Fagron stock. I have not been provided the information necessary to calculate damages for these warrants. I reserve the right to make any corrections or additions to my report based upon any new or additional information, documents or materials that become available.

## Summary of Alleged Damages

	<b>Alleged Damages</b>
<b>1. <u>Method 1: Based on Stock Price Inflation</u></b>	
A. Price Decline Only Through 8/6/15 Using Event Study	\$2.51M–\$2.68M
B. Using Price Decline due to Reimbursement Changes According to Analyst	\$3.88M–\$4.15M
C. Using Reduction in Analysts' Price Targets	\$3.25M
<b>2. <u>Method 2: Based on Standalone Value of AnazaoHealth</u></b>	<b>\$4.13M</b>

### ***Method 1: Damages based on inflation in stock price due to alleged misrepresentations and omissions***

13. For the first method, I measured damages as the difference in value between what Mr. Beckel believed he was receiving for the sale of AnazaoHealth, and what he actually received because of Defendants' misrepresentations and omissions. This method involved estimating the amount by which the value of the Fagron shares received by Mr. Beckel was inflated by the alleged misrepresentations and omissions. To estimate the amount of inflation in the stock, I analyzed how the stock price and the market reacted when the information that Defendants failed to disclose about Fagron's involvement with billing practices regarding compound pain creams, and about the negative impact of the changes in insurers' reimbursement policies on Fagron's business, was publicly announced.

14. To identify when the information that Defendants failed to disclose about Fagron's involvement with billing practices regarding compound pain creams, and about the negative impact of the changes in insurers' reimbursement policies on Fagron's business, was announced to the market, I reviewed publicly available information related to Fagron, TRICARE and the pharmaceutical industry, including Fagron's press releases and conference calls, news stories from Factiva and Bloomberg, L.P. and analyst reports on Fagron.<sup>17</sup>

<sup>17</sup> Bloomberg, L.P. is a commonly used provider of financial data and news. Factiva is an online reporting service and archive owned by Dow Jones & Company, Inc. that aggregates news content from nearly 33,000 sources from around the world.



15. Analyst reports are periodic reports issued by professional financial analysts at brokerage firms who perform research and analysis on specific industries and companies. Analysts analyze companies by studying publicly available information, such as SEC filings, as well as participating on conference calls and attending investor conferences where they can ask questions directly to management. Analysts use this information to model and value companies and industries using financial techniques such as discounted cash flow models and valuation multiples. Using these valuations, analysts typically issue price targets (*i.e.*, what price they expect the stock of a company to be in a certain time period), provide estimates reflecting their expectations of the company's future financial performance (such as estimates of future revenue, profits and earnings per share ("EPS")), and give recommendations to buy, hold or sell the stock. Analysts typically issue reports after new information about the company is released. These reports play an important role in disseminating information about a stock and can be a valuable source of information on market knowledge and sentiment at the time. There were at least 18 analyst reports on Fagron by 4 different analysts issued in 2015 and 2016.<sup>18</sup>

16. Based on my review, I find that the information that Defendants failed to disclose about Fagron's involvement with billing practices regarding compound pain creams, and about the negative impact of the changes in insurers' reimbursement policies on Fagron's business, was announced to the market on August 4, 2015.

17. On August 4, 2015, Fagron announced its financial results for the first half of FY2015 (*i.e.*, 1H15) and reported total revenue of €243.8 million, an increase of 16.6% over the prior year, and REBITDA of €65.6 million, an increase of 17.9% over the prior year. The table below shows Fagron's revenue by segment and REBITDA for 1H15 and 1H14, along with the company's reported year-over-year growth and organic growth (*i.e.*, growth excluding the impact of acquisitions):

---

<sup>18</sup> Based on available analyst reports from Thomson Reuters.

<b>Fagron Reported Revenue, REBITDA and Growth</b> <b>1H15 vs. 1H14</b>				
<b>Segment</b>	<b>1H15</b>	<b>1H14</b>	<b>Y-o-Y Growth</b>	<b>Organic Growth<sup>1</sup></b>
<b>All Segments</b>	<b>€243.8M</b>	<b>€209.1M</b>	<b>+16.6%</b>	<b>+2.6%</b>
Services	€92.2M	€57.4M	+60.7%	+14.8%
Essentials	€120.5M	€124.1M	-2.9%	-5.8%
Trademarks	€25.6M	€22.7M	+12.7%	+11.0%
HL Technology	€5.5M	€5.0M	+9.6%	-5.2%
<b>REBITDA</b>	<b>€65.6M</b>	<b>€55.7M</b>	<b>+17.9%</b>	n/a
<b>Notes and Sources:</b>				
Fagron 1H15 earnings press release: "Turnover grew 16.6% to € 243.8 million," August 4, 2015.				
<sup>1</sup> In constant exchange rates.				

18. Although Fagron did not report the breakdown of its 1H15 results in terms of 1Q15 and 2Q15 figures, analysts covering the company identified a sharp slowdown in Fagron's 2Q15 organic growth by comparing the 1H15 results to the company's previously reported 1Q15 results (announced in April 2015). For example, the ABN Amro analysts covering Fagron included the following data in a report on the company published on August 14, 2015:

### Fagron Organic Growth in 1Q15, 2Q15 and 1H15

According to ABN Amro

Segment	1H15 Revenue	Organic Growth <sup>1</sup>		
		Reported 1Q15	Reported 1H15	Implied 2Q15
<b>All Segments</b>	<b>€243.8M</b>	<b>+10.2%</b>	<b>+2.6%</b>	<b>-3.2%</b>
Services	€92.2M	+32.1%	+14.8%	+5.7%
Essentials	€120.5M	+0.2%	-5.8%	-11.0%
Trademarks	€25.6M	+4.6%	+11.0%	+16.7%
HL Technology	€5.5M	+0.1%	-5.2%	-10.4%

**Notes and Sources:**

"Where is the growth?" *ABN Amro*, August 14, 2015.

<sup>1</sup> In constant exchange rates.

19. Analysts covering Fagron attributed the sharp slowdown in 2Q15 growth at Fagron's two largest segments, the Services segment (from 32.1% growth in 1Q15 to 5.7% in 2Q15) and the Essentials segment (from 0.2% growth in 1Q15 to -11% in 2Q15), to changes in the reimbursement system in the US implemented in May/June 2015. Statements by all four analysts that released reports on Fagron following the 1H15 earnings announcement are shown below.

- a) KBC Securities: The KBC analysts stated that Fagron's 1H15 results represented the "weakest" half-year of organic growth *ever* reported by the company, and attributed these results to the "changing and challenging environment in the US":

Fagron's 1H15 financials disappointed as the top-line fell short of expectations due to a weak 2Q15 which was impacted by the changing and challenging environment in the US. [...]

**1H15 HAD WEAKEST ORGANIC SALES GROWTH EVER**

2Q15 was weakest ever quarter based on organic growth[.] Fagron's 1H15 organic growth of 3% at constant exchange rates was the company's weakest ever half-year organic growth. The disappointing result was entirely driven by 2Q15, which reported a 3% organic sales decline. [KBC Securities, 8/14/15, emphasis original]

The KBC analysts specified that changes in TRICARE's coverage policies for compounded medication were responsible for Fagron's disappointing 1H15 results:

Topical creams to treat pain accounted for the biggest increases in compound drugs costs in recent years, rising from \$ 5m in 2004 to \$ 514m in 2014 to more than \$ 500m in the month of April alone.

However, as of May 1 2015, Tricare began screening all ingredients in compound drugs to ensure they are FDA-approved and covered by Tricare. Screening was further tightened mid-May to address the continued influx of claims for compounds of what they call "dubious clinical evidence and excessive cost".

This action has had a massive impact. In April, Tricare spent \$ 18m a day in compound drug claims. By July, the daily costs had dropped to \$ 360,000 as a result of newly implemented screening processes and a crackdown on fraudulent prescriptions. With the new administrative controls, the average cost of compound prescriptions fell from \$ 6,889 in April to \$ 200. This pricing level is in-line with what Fagron disclosed during the 1H15 update. [...]

Apparently Tricare's decision to apply stricter rules on reimbursing compounded products as of May and the blacklisting of certain pharmacies have had a huge impact on the prescription market. Fagron's strong 1Q15 and weak 2Q15 are a perfect reflection of Tricare's 1Q and 2Q observations. [KBC Securities, 8/14/15]

- b) ABN Amro: The ABN Amro analysts stated that the primary factor causing the "alarming" slowdown in organic growth were the "reimbursement cuts in the US."

Fagron reported 1H15 results below both ABNe and consensus in terms of revenue. The miss was driven by an alarming decline in organic growth, resulting from a changed reimbursement system in the US. [...] The most negative factor impacting revenue in 1H15 was the slowdown in organic growth, mainly caused by reimbursement cuts in the US. [...] This slowdown in organic growth is due to a changed reimbursement system in the US, leading to lower prices for compounded medication and a decline in the Fagron Essentials. [ABN Amro, 8/14/15]

- c) ING: The ING analyst stated that the news of the impact of the changes in the US reimbursement system on Fagron's business was "surprising" given Fagron's prior statements on this issue – in particular, Fagron's prior statements that it did not expect other insurers, such as TRICARE, to change their reimbursement policies:

The miss on sales was driven by a marked slowdown of 2Q15 organic growth at CER [constant exchange rates] for Fagron towards -3.4% (1Q15: 10.5%), on the back of: (1) the phasing out of non-core low margin business (€m or 1.7ppt of 2Q14 turnover); and (2) more importantly, the impact of changes in the US reimbursement system for compounded drugs. The slowdown is more surprising as, during the 1Q15 conference call, management stated that it did not anticipate any additional negative impact of additional insurers moving into the compound management plan of Express Scripts (ie, Tricare, from May 2015 onwards). [ING, 8/6/15]

- d) Jefferies: The Jefferies analysts stated that the impact of “pricing headwinds” on the Essentials segment and the decline in the average price of Fagron’s non-sterile portfolio in the Services segment (which included compound pain creams) would likely cause the market to “discount” the company’s progress in sterile medication, and lead to “share price downside”:

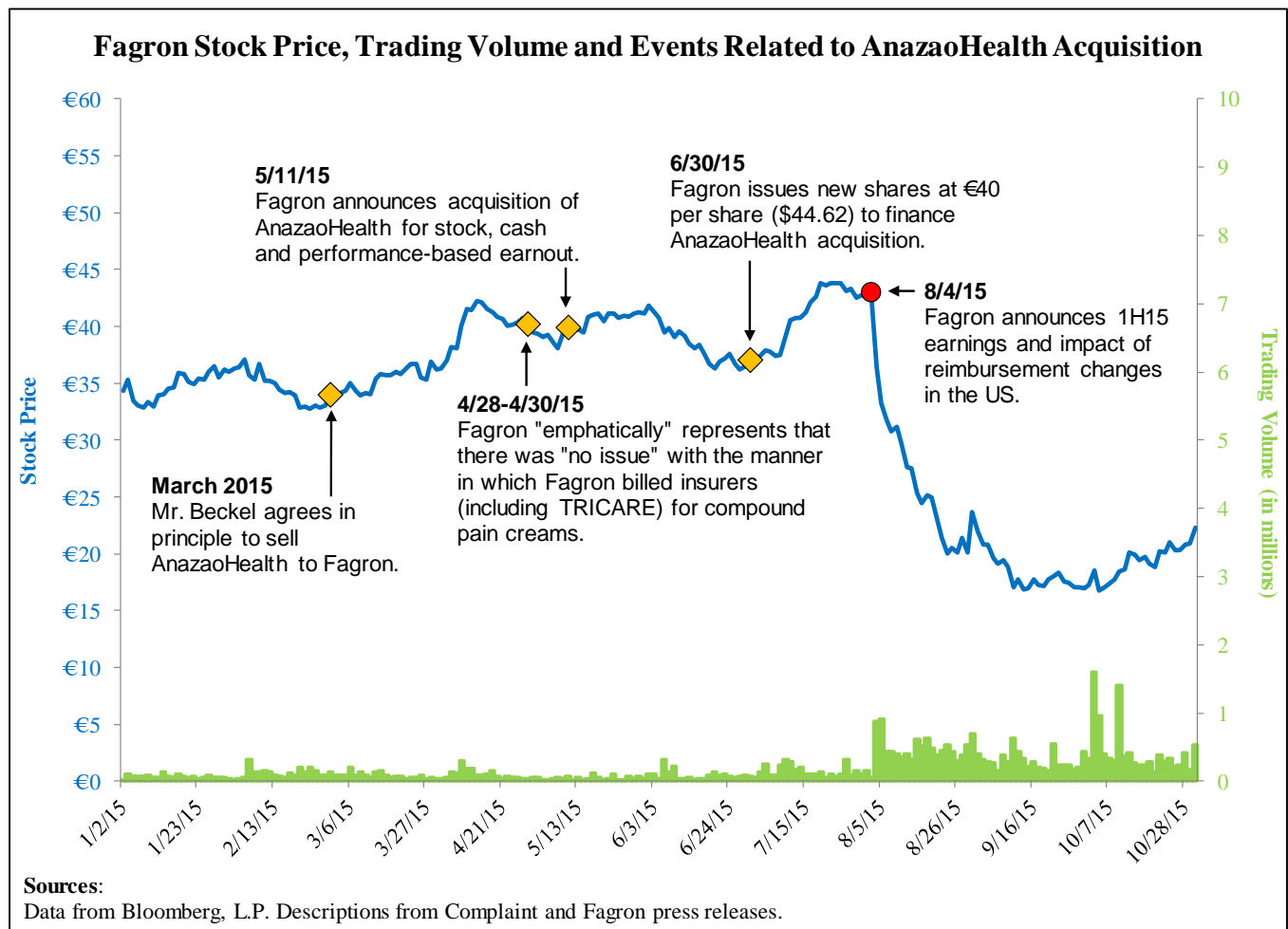
Fagron reiterated guidance despite missing cons and suffering c.€4m of pricing headwinds just since May, as the av. price of its non-sterile portfolio fell from c.\$275-300 to \$180-200m. The US strategy is likely to be discounted despite progress in sterile compounding in our view, which we believe could even result in short-term share price downside to c.€20 @c.11x '16E P/E; an absolute level not seen since prior to non-core disposals and the buy-in to US consolidation. [Jefferies, 8/17/15]

20. Later public filings by Fagron confirm that the changes in the reimbursement system in the US in FY2015 had impacted Fagron’s financial performance. For example, in a prospectus issued in June 2016, Fagron stated that the changes in the reimbursement system had resulted in “sharp declines in its sales of non-sterile compounded medication in 2015” at the Services segment (which operated in the US through the subsidiary Bellevue Pharmacy) and had led to a “significant” reduction in purchases of pharmaceutical raw materials at the Essentials segment (which operated in the US through the subsidiary Freedom Pharmaceuticals):

As a result of changes to the reimbursement regime for non-sterile compounding in the US implemented in 2015, Bellevue Pharmacy experienced sharp declines in its sales of non-sterile compounded medication in 2015 (resulting as of January 2016 in a classification as discontinued operations and a stop of production in March 2016). The changes to the reimbursement regime further resulted in the Group's pharmacy customers in the US significantly reducing their purchases of APIs [active pharmaceutical ingredients], primarily impacting Freedom Pharmaceuticals.<sup>19</sup>

<sup>19</sup> Fagron Prospectus, dated June 15, 2016, p. 18.

21. The chart below shows Fagron's stock price, trading volume, events related to the Acquisition and the announcement of the reimbursement changes:



22. I analyzed how the stock price and the market reacted when the information that Defendants failed to disclose about Fagron's involvement with billing practices regarding compound pain creams, and about the negative impact of the changes in insurers' reimbursement policies on Fagron's business, was publicly announced. To obtain an estimate of the price reaction due to the announcement of the reimbursement changes, I used the event study methodology.

23. An event study is a commonly accepted statistical analysis that measures the movement in a stock's price after an event or public announcement, typically adjusting for the

movement in the overall market and/or industry.<sup>20</sup> Academics often use an event study to determine how stock prices respond to new information.<sup>21</sup> An event study typically uses a statistical analysis called a regression to estimate the relationship between the company's daily stock returns and the daily returns of market and/or industry indices.<sup>22</sup> Using the regression results and the returns of the indices, the predicted stock price movement and abnormal stock price movement (or the amount the stock price moves in excess of the predicted amount) can be calculated for the day or period being tested. Then, the statistical significance of the abnormal stock price movement can be tested.<sup>23</sup>

24. The BEL 20 Index, the benchmark stock index for the Euronext Brussels stock exchange, was used to control for market movements. A control period of 252 trading days (approximately one year) before August 4, 2015 was used to run the regression.

25. According to the event study model, there was a statistically significant decline in Fagron's stock price on August 4, 2015. The statistically significant price reaction following the August 4 announcement continued for two additional days, until August 6, 2015. In total, the price decline through August 6 based on the event study was 26.5%. The results of the event study model following the announcement of the reimbursement changes are shown in the table below:

---

<sup>20</sup> See, for example, Alexander, Janet C., "The Value of Bad News in Securities Class Actions," *UCLA Law Review* 41: 1994, Fischel, Daniel R., "Use of Modern Finance Theory in Securities Fraud Cases Involving Actively Traded Securities," *The Business Lawyer* 38: 1982, and Dunbar, Frederick C., and David I. Tabak, "Materiality and Magnitude: Event Studies in the Courtroom," *Litigation Services Handbook: The Role of the Financial Expert* (John Wiley & Sons, Inc.: New York, NY, 3<sup>rd</sup> ed., 2001), ch. 19.

<sup>21</sup> See, for example, MacKinlay, A. Craig, "Event Studies in Economics and Finance," *Journal of Economic Literature*, 35: 1997, and Bowman, Robert G., "Understanding and Conducting Event Studies," *Journal of Business Finance & Accounting* 10(4): 1983.

<sup>22</sup> Regression analysis is used to estimate the relationship between two or more variables. See, for example, Hogg, Robert V. and Elliot A. Tanis, *Probability and Statistical Inference* (Prentice Hall: Upper Saddle River, NJ, 5<sup>th</sup> ed., 1997).

<sup>23</sup> The results of the event study are based on the 5% significance level, the standard typically used. See, for example, Freedman, David A., and David H. Kaye, "Reference Guide on Statistics," *Reference Manual on Scientific Evidence* (Washington, D.C.: The National Academies Press, 3<sup>rd</sup> ed., 2011), pp. 211-302, and Fisher, Franklin M., "Multiple Regression in Legal Proceedings," *Columbia Law Review*, 80: 1980.



**Price Reaction following Announcement of Reimbursement Changes**  
Controlling for the BEL 20 Index

<b>Date</b>	<b>Fagron Closing Price</b>	<b>Fagron Trading Volume</b>	<b>Fagron Return</b>	<b>BEL 20 Index Return</b>	<b>% Price Reaction</b>	<b>t-statistic</b>	<b>Stat. Sig.?<sup>1</sup></b>
8/3/15	€42.98	62,873					
<b>8/4/15</b>	<b>€36.40</b>	<b>876,891</b>	<b>-15.3%</b>	<b>0.0%</b>	<b>-15.3%</b>	<b>-7.14</b>	<b>Yes</b>
8/5/15	€33.30	904,720	-8.5%	0.8%	-9.0%	-4.21	Yes
8/6/15	€31.82	433,459	-4.5%	0.3%	-4.6%	-2.15	Yes
<b>Sum of Price Reactions</b>					<b>-26.5%</b>		

**Notes and Sources:**

Data from Bloomberg, L.P. Returns are predicted using the daily percent returns of Fagron as a function of the daily percent returns of the BEL 20 Index, regressed over a control period of 252 trading days (approximately one year) prior to August 4, 2015.

<sup>1</sup> Significance is based on the price reaction's t-statistic, calculated as the price reaction divided by the standard error of the regression over the sample period. "Yes" indicates significance at the 5% level.

26. Using the 26.5% stock price decline based on the event study yields damages of \$2.68 million ( $\$10.11 \text{ million} \times 26.5\% = \$2.68 \text{ million}$ ).

27. As an alternative, I reviewed analyst reports issued after the August 4, 2015 to determine how much analysts believed Fagron's stock price reacted to the announcement of the reimbursement changes. The Jefferies analysts covering Fagron stated that Fagron's stock price had declined by 41% following the announcement of the reimbursement changes as Fagron management's "credibility" had been "dented" after Fagron reiterated guidance, despite the reimbursement changes that had caused prices for Fagron's non-sterile compounded medication (which included compound pain creams) to decline "significantly":

Fagron's shares have fallen by 41% since 1H, as we believe management credibility has been dented by reiterating guidance while the pricing of its US non-sterile portfolio has declined significantly; we are now below FY guidance. We also believe insider buying is needed to signal value to the market since some management selling post 1Q, so we see limited share price support on fundamentals until visibility improves. [Jefferies, 8/17/15]

28. Using the 41% stock price decline in the Jefferies analyst report as a measure of inflation yields damages of \$4.1 million ( $\$10.11 \text{ million} \times 41\% = \$4.15 \text{ million}$ ).

29. Note that, in addition to announcing that the Essentials segment had been impacted by changes in the reimbursement system in the US, Fagron announced that, in 2Q15, the company had begun a project to phase out “non-strategic, low margin products.” According to Fagron, the phase-out of these products was expected to reduce the Essentials segment’s revenues but improve the segment’s profitability:

In the second quarter of 2015, a project to optimise the product portfolio and production process was started. This resulted in the phase-out of non-strategic, low-margin products. In the second quarter, the impact on turnover was approximately €2 million. For the full year, we estimate that the impact on turnover of our decision to phase out low-margin products will be between €12 and 15 million. So the impact of this decision on turnover will be negative, but we expect that it will have a positive impact on profitability.<sup>24</sup>

30. Analysts similarly stated that the phase-out of low margin products would positively impact the Essentials segment’s profitability. For example, the KBC analysts covering Fagron stated that, while the phase-out would have a negative impact on revenue “in the short term,” the impact on profitability would be “substantial” going forward, and noted that the last time Fagron had performed a similar phase-out, the Essentials segment’s revenues still grew:

As part of its strategic review, more low-margin products, usually with a low turnover ratio, will be phased out in 2015. Although this will have a negative impact on turnover in the short term, the impact on profitability as a percentage of sales and on the required working capital is said to be substantial. [...]

In 2014, Fagron phased out low margin products worth €11m, but still allowed the Essentials business to grow 5% organically (including the phased-out sales, Essentials’ organic growth was 9% in FY14). [KBC Securities, 8/14/15]

31. I estimated damages excluding any potential negative reaction to the “short-term” decline in revenue caused by the phase-out of low-margin products at the essentials segment. According to the ABN Amro analysts covering the company, the phase-out was expected to account for up to a €20 million reduction in the Essentials segment’s revenue in FY2015 and FY2016 – 50% of the total reduction in revenue in that period (with the other 50% due to the reimbursement changes):

---

<sup>24</sup> Fagron 1H15 earnings conference call, August 4, 2015.

In all, without much compensation from increased volumes, Essentials stands to lose some EUR 20m in revenue due to the PBM changes and an additional EUR 20m from product pruning in the next 12 months. [ABN Amro, 8/14/15]

32. Applying this 50% proportion to the change in analysts' estimates of the Essentials segment's future revenue for FY2015 and FY2016 from before to after the announcement of the reimbursement changes shows that only 6.5% of the reduction in Fagron's future revenue is attributable to the phase-out of low margin products at the Essentials segment:

<b>Analyst Estimates of Fagron Total Revenue Before and After Announcement of Reimbursement Changes</b>					
<b>Estimate Year</b>	<b>Total Fagron Revenue</b>			<b>Amount Attributed to</b>	
	<b>Before Annct.</b>	<b>After Annct.</b>	<b>Change</b>	<b>Essentials Segment</b>	<b>Product Phase-Out</b>
2015	€513.6M	€483.6M	-€30.0M	-€12.2M	<b>-€6.1M</b>
2016	€556.2M	€508.4M	-€47.8M	-€19.6M	<b>-€9.8M</b>
2017	€596.9M	€545.3M	-€51.6M	-€20.0M	-
2018	€639.2M	€583.7M	-€55.5M	-€20.4M	-
2019	€682.6M	€623.2M	-€59.4M	-€20.9M	-
		<b>Total</b>	<b>-€244.3M</b>	<b>-€93.1M</b>	<b>-€15.9M</b>
<b>Sources:</b>					
"Where is the growth?" <i>ABN Amro</i> , August 14, 2015, "Solid 1Q; US Deal Boosts Sterile Capabilities," <i>Jefferies</i> , May 15, 2015, and "A Setback to Progress," <i>Jefferies</i> , August 17, 2015.					

As the table shows, analysts reduced Fagron's future revenue (FY2015-FY2019) by a total of €244.3 million. Of the €244.3 million reduction, €93.1 million (or 38%) was due to the Essentials segment. Applying analysts' estimate that 50% of the reduction in Essentials segment revenue in FY2015 and FY2016 was due to the phase-out of low margin products yields a total revenue reduction of €15.9 million due to the product phase-out (a €6.1 million reduction in FY2015 plus a €9.8 million reduction in FY2016). The €15.9 million reduction in Essentials segment revenue accounts for just 6.5% of the €244.3 million reduction in total revenue.

33. Excluding the 6.5% portion of the price decline following the August 4, 2015 announcement that could potentially be due to the low-margin product phase-out yields damages of \$2.51 million using the price reaction based on the event study ( $\$2.68 \text{ million} \times 93.5\%$ ), and \$3.88 million using the stock price decline from the Jefferies August 17 analyst report ( $\$4.15 \text{ million} \times 93.5\%$ ).

34. Along with analyzing the stock price decline, I analyzed the changes in contemporaneous valuations of Fagron's stock by analysts covering the company following the announcement of the information that Defendants failed to disclose about Fagron's involvement with billing practices regarding compound pain creams and about the negative impact of the changes in insurers' reimbursement policies on Fagron's business.

35. As discussed above, analysts perform research and analysis on specific industries and companies, and analyze publicly available information, such as public filings, earnings releases and information from management from conference calls and investor conferences, to model and value companies and industries using financial techniques such as discounted cash flow models and valuation multiples. Using these valuations, analysts typically issue price targets (*i.e.*, what price they expect the stock of a company to be in a certain time period).

36. A review of the analyst reports issued after Fagron's 1H15 earnings announcement shows that all four analysts covering the company reduced their price targets for Fagron stock by an average of 32%.<sup>25</sup> The changes in analysts' price targets are summarized in the table below:

---

<sup>25</sup> Note that three of the four analysts also reduced their investment ratings for Fagron stock (*i.e.*, whether they recommend that investors buy, hold, or sell the stock) from "Buy" to "Hold." The analysts from Jefferies had already reduced their investment rating for Fagron stock to "Hold" prior to August 4, 2015.

<b>Change in Analysts' Price Targets Following Announcement of Reimbursement Changes</b>				
<b>Analyst</b>	<b>Price Target</b>			
	<b>Old</b>	<b>New</b>	<b>Change</b>	
ING	€46.50	€34.00	-€12.50	-27%
ABN Amro	€43.00	€29.00	-€14.00	-33%
KBC Securities	€48.50	€32.00	-€16.50	-34%
Jefferies	€40.00	€26.00	-€14.00	-35%
			<b>Average</b>	<b>-32%</b>
<b>Source:</b> Based on available analyst reports from Thomson Reuters.				

37. All four analysts that issued reports on Fagron following the 1H15 earnings announcement attributed the reduction in their price targets and/or investment ratings either directly to the changes in the reimbursement system in the US, or to the uncertainty/lack of visibility around Fagron's future financial prospects given the reimbursement changes, while *none* of these analysts stated that they were decreasing the price targets or investment ratings in response to the phase-out of low margin products at the Essentials segment:

- a) ABN Amro: The ABN Amro analysts attributed the lowered price target and investment rating for Fagron stock to the "uncertainty around the changing reimbursement system" and the negative impact of this uncertainty on the Services segment (via script prices for non-sterile compounded medication) and the Essentials segment:

The uncertainty around the changing reimbursement system and the negative impact this has on Fagron's script prices and Fagron Essentials concerns us. Although management remains positive about the US developments and believes in volume growth to compensate for the price decline, we have become more cautious after the disappointing results of 2Q15 and because concrete evidence that the situation is truly improving

is missing. The fact that we have no comparable numbers due to a recent change in reporting does not help in limiting our caution.

We continue to see strong growth potential in the compounding market for Fagron, especially in the sterile segment. However, so far we have only seen a negative impact in the US and we have become less confident about partly basing our assumptions on company guidance. We take a more cautious view and want to see proof of volume compensating for lower prices before assuming higher growth. Furthermore the company is trading dangerously close to its covenants which makes the room for error limited and additional M&A will need to be financed by an equity issue. As a result, we change our recommendation and target price and move to a Hold rating with a new target price of EUR 29. [ABN Amro, 8/14/15]

- b) KBC Securities: The KBC analysts attributed the lowered price target and investment rating to the “stressed” US reimbursement situation, its “heavy impact” on Fagron’s 2Q15 results, and the “increasing uncertainty” caused by the reimbursement changes:

**1H15 UPDATE PROMPTED DOWNWARD REVISION OF OUR FORECASTS**

Following the stressed US situation and its heavy impact on 2Q15, we have revised our FY15 and FY16 turnover forecasts by -4% and -6%, and net profit by ~25%. The high leverage is hampering the buy & build strategy and denting growth ambitions. To sustain this strategy and to meet the 2016 turnover objectives, Fagron will have to find alternative sources of funding, other than the cheap debt used so far.

A DCF-derived assessment points to a fair value of €32/sh, assuming mid-single-digit top-line growth and a sustainable 26% EBITDA margin. Based on our 2016 EBITDA assumption, the group has at our target price an 10x EV/EBITDA16 and 17x PE16 multiple. However, because of the increasing uncertainty on the US market, we recommend caution and adopt a neutral (HOLD) position, even after the strong share price correction. [KBC Securities, 8/14/15, emphasis original]

- c) Jefferies: The Jefferies analysts believed Fagron would trade at a “significant” discount to other companies in its industry due to “market concerns” around the reimbursement changes in the US:

**We believe a significant sector discount is warranted:** We expect Fagron’s shares to detach from fundamentals for some time, as see limited support with market concerns propagating around the US, uncertainties surrounding guidance and management’s share trading activity; post 1Q management sold c.€5.8m stock. We lower EPS by 17%-19%. [...]

**Valuation:** Our new €26 PT is derived by applying a 14.0x P/E multiple in 2016E as we expect the shares to trade at a discount to fundamentals. [Jefferies, 8/17/15, emphasis original]

- d) ING: The ING analyst attributed the lowered price target and investment rating for Fagron stock to the “lower visibility and lower organic growth” from the “impact of the changed US reimbursement system”:

**We downgrade from Buy to HOLD** as we believe that the risk/reward profile of the shares have further weakened as: (1) acquisitions will, in our view, be financed by issuance of equity; and (2) there is less visibility on the impact of the changed US reimbursement system. [...]

**We lower our TP [target price] to €34**, valuing Fagron at 10.5x 2016F EV/REBITDA [enterprise value/REBITDA], down from 12x previously, given the lower visibility and lower organic growth. We furthermore remove the redeployment potential from our TP as it seems increasingly likely that new acquisitions will be paid for through the raising of equity. [ING, 8/6/15, emphasis original]

38. Using the 32% decrease in analysts’ price targets as a measure of inflation yields damages of \$3.25 million ( $\$10.11 \text{ million} \times 32\% = \$3.25 \text{ million}$ ).

***Method 2: Damages based on standalone value of AnazaoHealth if it had not been sold to Fagron***

39. For the second method, I estimated damages as the difference between what Mr. Beckel would have had today had he been told the information about Fagron’s involvement with billing practices that Defendants failed to disclose and had not sold AnazaoHealth to Defendants, and what he actually received. This method entailed estimating how the value of AnazaoHealth would have changed between the time of the Acquisition and today if it had not been acquired and had instead remained a standalone company.

40. I estimated the difference between what Mr. Beckel has today in Fagron stock and what he would have had if AnazaoHealth’s value moved with the industry. It is my understanding that Mr. Beckel has sold approximately half of the 226,700 Fagron shares that he received in connection with the AnazaoHealth acquisition (for proceeds worth \$1.63 million), and that he continues to hold the remaining shares as of today (currently worth \$1.90 million). These figures are shown in the table below:

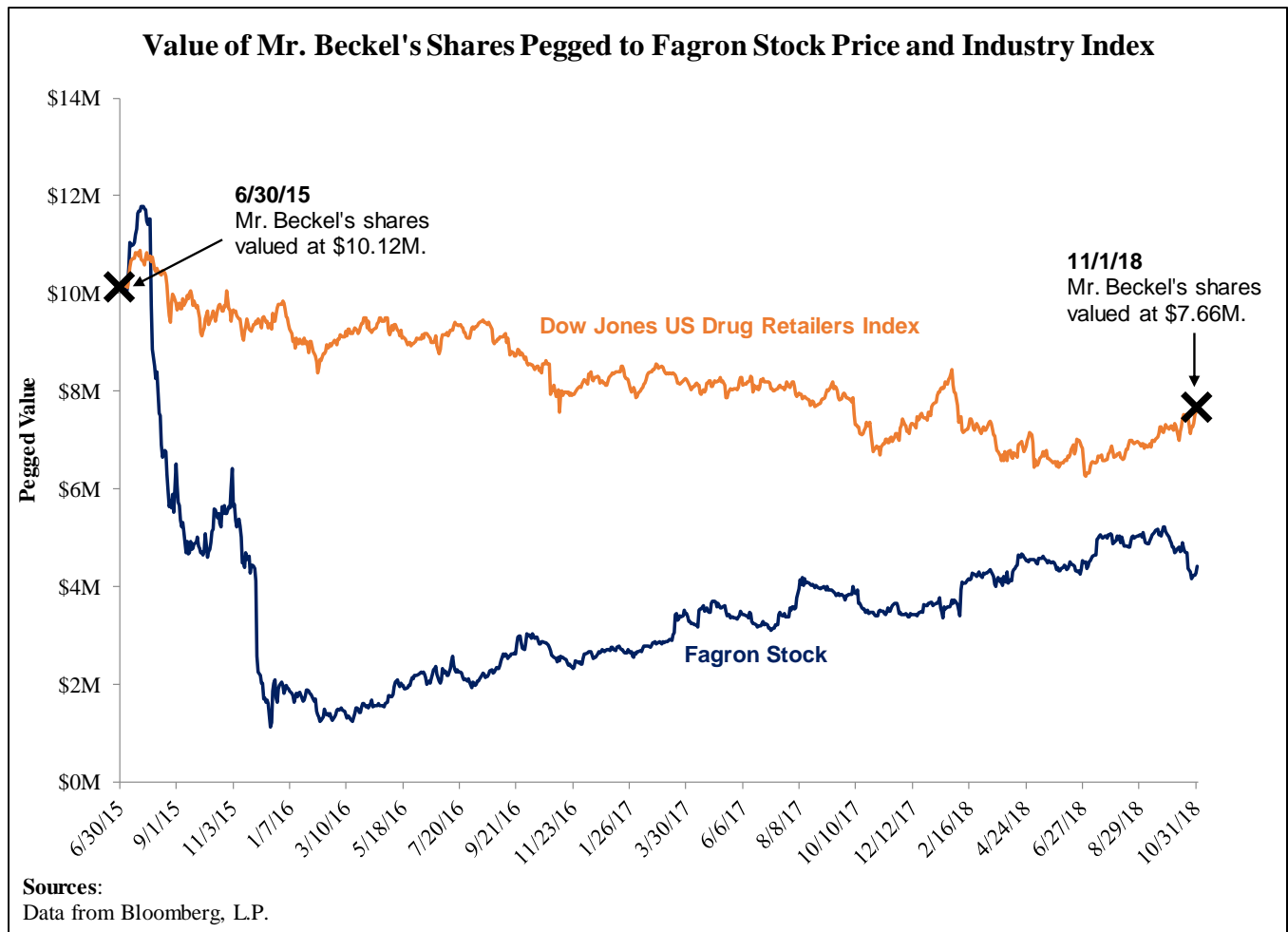


<b>Value of Mr. Beckel's Fagron Shares As of November 1, 2018</b>			
<b>Date</b>	<b># of Shares</b>	<b>Price</b>	<b>Proceeds</b>
<i>Share Sales</i>			
8/18/16	28,338	\$8.35	\$0.24M
9/9/16	28,337	\$9.78	\$0.28M
9/20/18	28,849	\$19.67	\$0.57M
9/21/18	27,827	\$19.60	\$0.55M
Total	113,351		\$1.63M <sup>1</sup>
<i>Unsold Shares</i>			
11/1/18	113,349	\$16.75 <sup>2</sup>	\$1.90M
<b>Total Value of Fagron Shares</b>			<b>\$3.53M</b>
<b>Notes and Sources:</b>			
"Beckel Family ARSUF History.pdf," Bloomberg, L.P. and the Federal Reserve Bank of St. Louis.			
<sup>1</sup> Total proceeds as of November 1, 2018, grown by the interest rate paid by 3-month US Treasury Bills.			
<sup>2</sup> Based on Fagron's closing price of €14.68 on November 1, 2018, converted to USD using the USD-EUR exchange rate of \$1 = €0.88 according to Bloomberg, L.P.			

41. I assumed that AnazaoHealth would have moved consistently with the drug retail industry, the industry that AnazaoHealth was classified under according to FactSet, a provider of financial data that caters to finance industry professionals.<sup>26</sup> To estimate how the value of AnazaoHealth would have changed, I used the Dow Jones US Drug Retailers Index.<sup>27</sup> As the chart below shows, Mr. Beckel's \$10.11 million in Fagron shares would have currently been worth \$7.66 million if he had not sold AnazaoHealth:

<sup>26</sup> The specific industry codes that AnazaoHealth was classified under are: NAICS code 446110 (Pharmacies and Drug Stores), SIC code 5912 (Drug Stores and Proprietary Stores) and FactSet industry code 3510 (Drugstore Chains).

<sup>27</sup> As an alternative, I performed the analysis assuming AnazaoHealth's value would have moved consistently with the pharmaceuticals industry. Using a benchmark index for the pharmaceuticals industry, the Standard and Poor's Pharmaceuticals Select Industry Index, I obtained similar results (*i.e.*, damages of \$4.19 million).



42. Thus, damages are the difference between what Mr. Beckel has today in Fagron stock (*i.e.*, \$3.53 million), and what he would have had if AnazaoHealth's value moved with the industry (*i.e.*, \$7.66 million).<sup>28</sup> This difference equals \$4.13 million in damages.<sup>29</sup>

*Lucy P. Allen*

Lucy P. Allen

<sup>28</sup> Note that these figures do not include the value of any salary or benefits that Mr. Beckel might have received had he not sold AnazaoHealth and remained employed by the company.

<sup>29</sup> Note that I have not been provided the information necessary to perform an independent valuation of AnazaoHealth as of today. I reserve the right to make any corrections or additions to my report based upon any new or additional information, documents or materials that become available.



Lucy P. Allen  
Managing Director

NERA Economic Consulting  
1166 Avenue of the Americas  
New York, New York 10036  
Tel: +1 212 345 5913 Fax: +1 212 345 4650  
lucy.allen@nera.com  
www.nera.com

## Appendix A

### MANAGING DIRECTOR

#### Education

##### **YALE UNIVERSITY**

M.Phil., Economics, 1990

M.A., Economics, 1989

M.B.A., 1986

##### **STANFORD UNIVERSITY**

A.B., Human Biology, 1981

#### Professional Experience

- |                        |  |
|------------------------|--|
| 1994-Present           | <p><b>National Economic Research Associates, Inc.</b><br/><u>Managing Director</u>. Responsible for economic analysis in the areas of securities, finance and environmental and tort economics.<br/><u>Senior Vice President (2003-2016)</u>.<br/><u>Vice President (1999-2003)</u>.<br/><u>Senior Consultant (1994-1999)</u>.</p>   |
| 1992-1993              | <p><b>Council of Economic Advisers, Executive Office of the President</b><br/><u>Staff Economist</u>. Provided economic analysis on regulatory and health care issues to Council Members and interagency groups. Shared responsibility for regulation and health care chapters of the <i>Economic Report of the President, 1993</i>. Working Group member of the President's National Health Care Reform Task Force.</p>   |
| 1986-1988<br>1983-1984 | <p><b>Ayers, Whitmore &amp; Company (General Management Consultants)</b><br/><u>Senior Associate</u>. Formulated marketing, organization, and overall business strategies including:<br/>Plan to improve profitability of chemical process equipment manufacturer.<br/>Merger analysis and integration plan of two equipment manufacturers.<br/>Evaluation of Korean competition to a U.S. manufacturer.<br/>Diagnostic survey for auto parts manufacturer on growth obstacles.<br/>Marketing plan to increase international market share for major accounting firm.</p> |

Summer 1985      **WNET/Channel Thirteen, Strategic Planning Department**  
Associate. Assisted in development of company's first long-term strategic plan. Analyzed relationship between programming and viewer support.

1981-1983      **Arthur Andersen & Company**  
Consultant. Designed, programmed and installed management information systems. Participated in redesign/conversion of New York State's accounting system. Developed municipal bond fund management system, successfully marketed to brokers. Participated in President's Private Sector Survey on Cost Control (Grace Commission). Designed customized tracking and accounting system for shipping company.

Teaching  
1989- 1992      **Teaching Fellow, Yale University**  
Honors Econometrics  
Intermediate Microeconomics  
Competitive Strategies  
Probability and Game Theory  
Marketing Strategy  
Economic Analysis

## **Publications, Speeches and Conference Papers**

"Snapshot of Recent Trends in Asbestos Litigation: 2018 Update," (co-author), NERA Report, 2018.

"Trends and the Economic Effect of Asbestos Bans and Decline in Asbestos Consumption and Production Worldwide," (co-author), *International Journal of Environmental Research and Public Health*, 15(3), 531, 2018.

"Snapshot of Recent Trends in Asbestos Litigation: 2017 Update," (co-author), NERA Report, 2017.

"Asbestos: Economic Assessment of Bans and Declining Production and Consumption," World Health Organization, 2017.

"Snapshot of Recent Trends in Asbestos Litigation: 2016 Update," (co-author), NERA Report, 2016.

"Economic Dimension and Societal Costs and Benefits of Banning Asbestos," presented at the World Health Organization, Regional Office for Europe conference, Assessing the Economic Costs of the Health Impacts of Environmental and Occupational Factors: The Economic Dimension of Asbestos, Bonn, Germany, 2016.

"Snapshot of Recent Trends in Asbestos Litigation: 2015 Update," (co-author), NERA Report, 2015.

Lucy P. Allen

Participant in panel on “Expert Reports and Depositions” at PLI Expert Witness 2014, hosted by the Practising Law Institute, New York, New York, 2014.

“Snapshot of Recent Trends in Asbestos Litigation: 2014 Update,” (co-author), NERA Report, 2014.

“High Frequency Trading --A Primer in 1,800,000 Milliseconds” before the Litigation Group at Morrison Foerster, New York, New York, 2014.

“Snapshot of Recent Trends in Asbestos Litigation: 2013 Update,” (co-author), NERA Report, 2013.

“Asbestos Payments per Resolved Claim Increased 75% in the Past Year – Is This Increase as Dramatic as it Sounds? Snapshot of Recent Trends in Asbestos Litigation: 2012 Update,” (co-author), NERA Report, 2012.

“Snapshot of Recent Trends in Asbestos Litigation: 2011 Update,” (co-author), NERA White Paper, 2011.

Participant in panel at The Implications of Matrixx, hosted by NERA Economic Consulting, New York, New York, 2011.

“2011 & Beyond–Predicting Mass Tort Litigation: with a Focus on Pharmaceutical Torts” presented at Emerging Insurance Coverage and Allocation Issues, hosted by Perrin Conferences, New York, New York, 2011.

Presented recent trends in settlements, predicting settlement amounts, and the use of economic analysis at mediation in the “Settlement Trends & Tactics” panel at Securities Litigation & Enforcement: Current Developments & Strategies, hosted by the New York City Bar, New York, New York, 2010.

“Snapshot of Recent Trends in Asbestos Litigation: 2010 Update,” (co-author), NERA White Paper, 2010.

“Settlement Trends and Tactics” presented at Securities Litigation During the Financial Crisis: Current Development & Strategies, hosted by the New York City Bar, New York, New York, 2009.

“GM and Chrysler Bankruptcies: Potential Impact on Other Asbestos Defendants” presented at Asbestos Litigation Conference: A Comprehensive National Overview and Outlook, hosted by Perrin Conferences, San Francisco, California, 2009.

“Snapshot of Recent Trends in Asbestos Litigation,” (co-author), NERA White Paper, 2009.

“Emerging Economies and Product Recall -- Are the Claims Coming?” presented at The International Reinsurance Summit 2008, Hamilton, Bermuda, 2008.

“China Product Recalls: What’s at Stake and What’s Next,” (co-author), NERA Working Paper, 2008.

“Recent Trends in Securities Litigation” presented at Strategies, Calculations & Insurance in Complex Business Litigation, hosted by the Directors Roundtable, New York, New York, 2008.

“The Current Landscape” presented at Mealey's Product Recall Liability Conference: Made in China and Beyond, Washington, DC, 2007.

“China Product Recalls: What's at Stake and What's Next” presented at China Product Recalls, sponsored by National Economic Research Associates, New York, New York, 2007.

“Damages and Loss Causation in Shareholder Class Actions after Dura” presented at Securities Litigation: Emerging Trends in Enforcement and Winning Litigation Strategies hosted by the International Quality & Productivity Center, New York, New York, 2006.

“Forecasting Product Liability by Understanding the Driving Forces,” (co-author), The International Comparative Legal Guide to Product Liability, 2006.

“Recent Trends in Securities Class Action Litigation,” presented at The Class Action Litigation Summit Program Class Action in the Securities Industry, Washington, D.C., 2003.

“Product Liability Claims Estimation – Four Steps, Four Myths” presented at Standard & Poor’s Seminar, New York, New York, 2001.

“How Bad Can It Be? The Economics of Damages and Settlements in Shareholder Class Actions,” Balancing Disclosure and Litigation Risks for Public Companies (Or Soon-To-Be Public Companies) Seminar, sponsored by Alston & Bird LLP and RR Donnelley Financial, Nashville, Tennessee, 2000.

“Securities Litigation Reform: Problems and Progress,” Viewpoint, November 1999, Issue No. 2 (co-authored).

“Trends in Securities Litigation and the Impact of the PSLRA,” Class Actions & Derivative Suits, American Bar Association Litigation Section, Vol. 9, No. 3, Summer 1999 (co-authored).

“Random Taxes, Random Claims,” Regulation, Winter 1997, pp. 6-7 (co-authored).

“Adverse Selection in the Market for Used Construction Equipment,” presented at the NBER Conference on Research in Income and Wealth, Federal Reserve Board, June 1992.

## **Expert Reports, Depositions & Testimony (4 years)**

Deposition Testimony and Expert Report before the United States District Court Northern District of Illinois Eastern Division in *In re the Allstate Corporation Securities Litigation*, 2018.

Expert Report before the United States District Court Central District of Californian Southern Division in *Steven Rupp et al. v. Xavier Becerra et al.*, 2018.

Rebuttal Report and Expert Report before the District Court for the State of Nevada in *Dan Schmidt v. Liberator Medical Holdings, Inc., et al.*, 2018.

Rebuttal Report and Expert Reports before the Clark County District Court of Nevada in *Round Square Company Limited v. Las Vegas Sands, Inc.*, 2018.

Supplemental Report and Expert Report before the United States District Court Middle District of Tennessee in *Zwick Partners LP and Aparna Rao v. Quorum Health Corporation, et al.*, 2018.

Declaration before the Superior Court of the State of Vermont in *Vermont Federation of Sportsmen's Club et al. v. Matthew Birmingham et al.*, 2018.

Deposition Testimony and Expert Report before the United States District Court Middle District of Tennessee in *Nikki Bollinger Grae v. Corrections Corporation of America et al.*, 2018.

Testimony and Expert Report before the American Arbitration Association in *Arctic Glacier U.S.A, Inc. and Arctic Glacier U.S.A., Inc. Savings and Retirement Plan v. Principal Life Insurance Company*, 2018.

Deposition Testimony and Expert Report before the United States District Court Southern District of New York in *Marvin Pearlstein v. Blackberry Limited et al.*, 2018.

Deposition Testimony and Expert Report before the United States District Court Eastern District of Texas in *Alan Hall and James DePalma v. Rent-A-Center, Inc., Robert D. Davis, and Guy J. Constant*, 2018.

Deposition Testimony, Surrebuttal Report, Rebuttal Report and Expert Report before the United States District Court Southern District of Iowa in *Mahaska Bottling Company, Inc., et al. v. PepsiCo, Inc. and Bottling Group, LLC*, 2018.

Testimony, Deposition Testimony and Declaration before the United States District Court District of New Jersey in *Association of New Jersey Rifle & Pistol Clubs, Inc. et al. v. Gurbir Grewal et al.*, 2018.



Lucy P. Allen

Deposition Testimony, Supplemental Report and Expert Report before the Supreme Court of the State of New York in *Bernstein Liebhard, LLP v. Sentinel Insurance Company, Ltd.*, 2018.

Expert Report before the District Court for Douglas County, Nebraska in *Union Pacific Railroad Company v. L.B. Foster Company and CXT Incorporated*, 2018.

Deposition Testimony and Declarations before the United States District Court Southern District of New York in *Andrew Meyer v. Concordia International Corp., et al.*, 2018.

Deposition Testimony before the United States District Court Southern District of California in *Virginia Duncan, et al. v. Xavier Becerra, et al.*, 2018.

Expert Report and Declaration before the United States District Court Southern District of California in *Virginia Duncan, et al. v. Xavier Becerra, et al.*, 2017.

Deposition Testimony and Expert Report before the United States District Court for the Western District of Texas, Austin Division in *City of Pontiac General Employees' Retirement System v. Dell, Inc., et al.*, 2017.

Deposition Testimony and Expert Report before the United States District Court for the Southern District of Texas, Houston Division in *In re Willbros Group, Inc. Securities Litigation*, 2017.

Declaration before the United States District Court Eastern District of California in *William Wiese, et al. v. Xavier Becerra, et al.*, 2017.

Deposition Testimony and Expert Report before the United States District Court for the Southern District of Texas, Houston Division in *In re Cobalt International Energy Inc. Securities Litigation.*, 2017.

Testimony, Deposition Testimony and Expert Report before the United States District Court for the Northern District of Texas, Dallas Division in *DEKA Investment GmbH, et al. v. Santander Consumer USA Holdings, Inc., et al.*, 2017.

Deposition Testimony before the Superior Court of the State of North Carolina for Mecklenburg County in *Next Advisor, Inc. v. LendingTree, Inc.*, 2017

Deposition Testimony and Expert Report before the Supreme Court of the State of New York, County of New York in *Iroquois Master Fund Ltd., et al. v. Hyperdynamics Corporation*, 2016.

Deposition Testimony and Expert Report before the United States District Court for the Northern District of Texas, Dallas Division in *The Archdiocese of Milwaukee Supporting Fund, Inc., et al. v. Halliburton Company, et al.*, 2016.

Lucy P. Allen

Expert Report before the United States District Court for the Northern District of Georgia, Atlanta Division, in *In re Suntrust Banks, Inc. ERISA Litigation*, 2016.

Deposition Testimony and Expert Report before the Superior Court of New Jersey, Union County, in *Syngenta Crop Protection, Inc. v. Insurance Company of North America et al.*, 2015.

Declaration before the United States District Court Northern District of Georgia, in *John Noble, et al. v. Premiere Global Services, Inc., et al.*, 2015.

Deposition Testimony and Expert Report before the United States District Court Central District of California, in *Amanda Sateriale, et al. v. RJ Reynolds Tobacco Co. et al.*, 2015.

Rebuttal Report and Expert Report in the United States of America before the Securities and Exchange Commission in *Houston American Energy Corp., et al.*, 2014.

Testimony, Deposition Testimony and Expert Report before the United States District Court for the Northern District of Texas, Dallas Division in *The Archdiocese of Milwaukee Supporting Fund, Inc., et al. v. Halliburton Company, et al.*, 2014.

Deposition Testimony and Expert Report before the United States District Court for the Eastern District of Pennsylvania in *Power Restoration International, Inc. v. PepsiCo, Inc., Bottling Group, LLC, and Frito-Lay Trading Company (Europe), GmbH*, 2014.

Deposition Testimony and Expert Reports before the United States District Court Southern District of New York in *In re Lower Manhattan Disaster Site Litigation*, 2014.

Deposition Testimony and Expert Report before the United States District Court Southern District of Florida in *Atul Kumar Sood, et al. v. Catalyst Pharmaceutical Partners Inc., et al.*, 2014.

Declaration before the Superior Court of Gwinnett County State of Georgia in *City of Riviera Beach General Employees Retirement System, et al. v. Aaron's Inc., et al.*, *Norfolk County Retirement System, et al. v. Aaron's Inc., et al.*, 2014.

Deposition Testimony, Surrebuttal Report and Expert Report before the United States District Court Middle District of Tennessee Nashville Division in *Garden City Employees' Retirement System and Central States, Southeast and Southwest Areas Pension Fund, et al. v. Psychiatric Solutions, Inc., et al.*, 2014.

Declaration before the United States District Court Northern District of California San Jose Division in *Fyock, et al. v. The City of Sunnyvale, et al.*, 2014.

Deposition Testimony and Expert Report before the United States District Court for the District of Maryland (Northern Division) in *Kolbe, et al. v. O'Malley, et al.*, 2014.

Lucy P. Allen

Declaration before the United States District Court Northern District of California in  
*San Francisco Veteran Police Officers Association, et al. v. The City and County of San  
Francisco, et al.*, 2014.

# Exhibit 71

VIX Volatility Products

# Turn Volatility to Your Advantage

Welcome to your go-to place for information about the VIX complex, including VIX [options](#) and [futures](#). Learn to measure, model and trade market moves with the world's widest array of volatility products and resources.

What is volatility?	▼
What is the VIX Index?	▼
How is the VIX Index calculated?	▼
How is the VIX Index used?	▼

Cboe

RMC

2023

Start Your Engines!

The Premier Risk Management Conference

Austin, TX | Oct 17 - 20, 2023

Learn more

Trade VIX Options Nearly 24 Hours a Day

[Learn More >](#)

## The VIX Index and Muted Volatility in 2022

[Read the article to learn more >](#)

# Making Sense of the VIX Index: An Indicator of Expected Market Volatility



2:53

Cboe's Inside Volatility Newsletter brings you the latest insights on the volatility market, breaking news, and interesting trades.

[Subscribe to Cboe's Inside Volatility Newsletter >](#)

VIX® Index Charts & Data



^VIX

13.66

-5.79%

Prev.Close

14.5

Open

14.49

52 Week

High 34.88

Low 13.5

as of June 16, 2023 at 2:34 PM EDT



Cboe is the home of volatility trading, and the Cboe Volatility Index® (VIX® Index) is the centerpiece of Cboe's volatility franchise, which includes VIX futures and VIX options.

- VIX Index

VIX Options

VIX Futures

Mini VIX Futures

## Overview

## VIX Methodology

The VIX Index is a calculation designed to produce a measure of constant, 30-day expected volatility of the U.S. stock market, derived from real-time, mid-quote prices of S&P 500® Index (SPX<sup>SM</sup>) call and put options. On a global basis, it is one of the most recognized measures of volatility -- widely reported by financial media and closely followed by a variety of market participants as a daily market indicator.

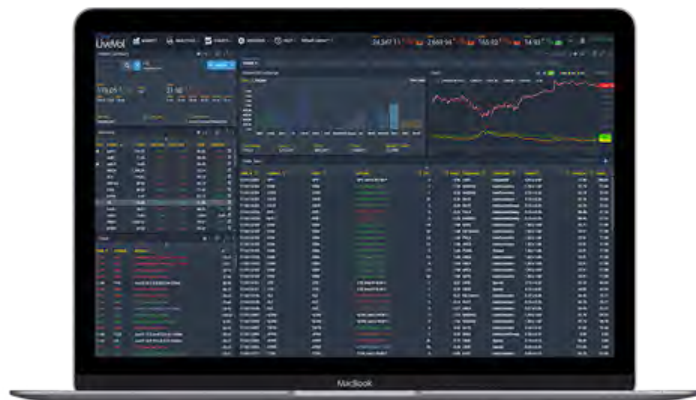
[Delayed Quotes >](#)

[Historical Data >](#)

# VIX Options Analytics

Get analysis on VIX Options and the rest of the U.S.-listed options market with Cboe LiveVol analytics platforms. LiveVol's web-based platforms provide everything you need to quickly analyze trading activity and identify opportunities.

View tutorials and take a free trial at [LiveVol.com](https://www.cboe.com/tradable_products/vix/).



# VIX<sup>®</sup> Futures & Options Strategies

VIX futures and options have unique characteristics and behave differently than other financial-based commodity or equity products. Understanding these traits and their implications is important. VIX futures and options may provide market participants with flexibility to hedge a portfolio, employ strategies in an effort to generate returns from relative pricing differences, or express a bullish, bearish or neutral outlook for broad market implied volatility.

## Portfolio Hedging

One of the biggest risks to an equity portfolio is a broad market decline. The VIX Index has had a historically strong inverse relationship with the S&P 500<sup>®</sup> Index. Consequently, a long exposure to volatility may offset an adverse impact of falling stock prices. Market participants should consider the time frame and characteristics associated with VIX futures and options to determine the utility of such a hedge.

## Long/Short Volatility

VIX futures provide a pure play on the level of expected volatility. Expressing a long or short sentiment may involve buying or selling VIX futures. Alternatively, VIX options may provide similar means to position a portfolio for potential increases or decreases in anticipated volatility.

## Risk Premium Yield

Over long periods, index options have tended to price in slightly more uncertainty than the market ultimately realizes. Specifically, the expected volatility implied by SPX option prices tends to trade at a premium relative to subsequent realized volatility in the S&P 500 Index. Market participants have used VIX futures and options to capitalize on this general difference between expected (implied) and realized (actual) volatility, and other types of volatility arbitrage strategies.

## Term Structure Trading

One of the unique properties of volatility – and the VIX Index – is that its level is expected to trend toward a long-term average over time, a property commonly known as "mean-reversion." The mean reverting nature of volatility is a key driver of the shape of the VIX futures term structure and the way it can move in response to changes in perceived risk. CFE lists nine standard (monthly) VIX futures contracts, and six weekly expirations in VIX futures. As such, there is a wide variety of potential calendar spreading opportunities depending on expectations for implied volatility.

### Term Structure Data and Charts >

The information above is provided for general education and information purposes only. No statement within these materials should be construed as a recommendation to buy or sell a security or future or to provide investment advice. Supporting documentation for any claims, comparisons, statistics or other technical data in these materials is available by contacting Cboe at [cboe.com/contact](https://www.cboe.com/contact).

# VIX<sup>®</sup> Index Research

[Visit Research Library >](#)

## S&P Dow Jones Indices: A Practitioner's Guide to Reading VIX

An easy-to-read guide for understanding the VIX complex. This document provides investors with simple guidelines that translate VIX Index levels into potentially more meaningful predictions or measures of market sentiment.

[Download Whitepaper >](#)

## BlackRock: VIX Your Portfolio

A research paper outlining the opportunities created by using market uncertainty. This paper explains how the strategy of selling volatility has generated higher returns with smaller losses, compared with traditional equity portfolios.

[Download Whitepaper >](#)

The inclusion of research not conducted or explicitly endorsed by Cboe should not be construed as an endorsement or indication of the value of any research.



© 2023 Cboe Exchange, Inc. All rights reserved.

### Company

[About Us](#)

[ESG at Cboe](#)

### Services

[Data and Access Solutions](#)

[Execution Services](#)

### Education

Careers

Investor Relations

Public Policy

Insights

Hours & Holidays

Locations

System Status

Contact Us

Listings

Access Services

European SI Services

European Trade Reporting Services

Research

RMC

Markets

U.S. Options

U.S. Equities

U.S. Futures

Canadian Equities

European Derivatives

European Equities

Foreign Exchange

Cboe Asia Pacific

Cboe Australia

Cboe Japan

Cboe Digital

Investor Relations

Cboe Press Releases

Investor Protection

Corp. Responsibility

---

Accessibility

Use of Content

Privacy Statement

Copyright, Trademark & Patents

Biometric Information Privacy Policy

OCC & Investor Protection

Terms&Conditions



# Exhibit 72



&lt; shut-in

x



Dictionary

Thesaurus

# shut-in

 1 of 3 noun

'shət-, in

[Synonyms of \*shut-in\* >](#)

- 1 : a person who is confined to home, a room, or bed because of illness or incapacity
- 2 : a narrow gorge-shaped part of an otherwise wide valley
- 3 : available oil or gas which is not being produced from an existing well

# shut-in

 2 of 3 adjective

'shət-'in

- 1 : confined to one's home or an institution by illness or incapacity
- 2 a : **SECRETIVE, BROODING**  
a bitter, *shut-in* face  
— Claudia Cassidy
- b : tending to avoid social contact : **WITHDRAWN**  
the *shut-in* personality type  
— S. K. Weinberg

# shut in

 3 of 3 verb

shut in; shutting in; shuts in

[transitive verb](#)



Dictionary

Thesaurus

## Recent Examples on the Web

---

### Noun

Caitlin Jett, a spokesperson for Newton County Sheriff's Office, declined to say whether another *shut-in* was planned last year.

— Faith Karimi, *CNN*, 30 Oct. 2023

Hurley also enjoyed visiting *shut-ins* and nursing home residents, entertaining the seniors with Irish tunes.

— Bob Goldsborough, *Chicago Tribune*, 12 Sep. 2023

[See More](#) ▾

These examples are programmatically compiled from various online sources to illustrate current usage of the word 'shut-in.' Any opinions expressed in the examples do not represent those of Merriam-Webster or its editors. [Send us feedback](#) about these examples.

## First Known Use

---

### Noun

1891, in the meaning defined at [sense 1](#)

### Adjective



Dictionary

Thesaurus

## Time Traveler

The first known use of *shut-in* was in the 14th century

[See more words from the same century](#)

[shut-eye](#)

**shut-in**

**shut in**

[See More Nearby Entries >](#)

Style

MLA

"Shut-in." *Merriam-Webster.com Dictionary*, Merriam-Webster, <https://www.merriam-webster.com/dictionary/shut-in>. Accessed 20 Nov. 2023.

 [Copy Citation](#)



# Exhibit 73

THE CBOE VOLATILITY INDEX®

# What Is VIX and What Does it Measure?

## The Industry Standard in Volatility Measurement and Forecasting

### S&P Dow Jones Indices

A Division of **S&P Global**



The Cboe Volatility Index, better known as VIX, projects the probable range of movement in the U.S. equity markets, above and below their current level, in the immediate future. Specifically, VIX measures the implied volatility of the S&P 500® (SPX) for the next 30 days. When implied volatility is high, the VIX level is high and the range of likely values is broad. When implied volatility is low, the VIX level is low and the range is narrow.

Since VIX reaches its highest levels when the stock market is most unsettled, the media tend to refer to VIX as a fear gauge. In the sense that VIX is a measure of sentiment—of worry in particular—the description is on the mark.

[English](#) ▾ [About](#) ▾ ▾ [Discover more](#) [Register](#) [Log in](#) [Feedback](#) [Help](#) [Contact](#) [S&P Global's offerings](#)

**S&P Dow Jones  
Indices**

A Division of **S&P Global**

**Indices**

**Research & Insights**

**Exchange Relationships**

**Professional Resources**

**Governance**



[English](#) ▾ [About](#) ▾ ▾ [Discover more](#)  [Get S&P Global's offerings](#)

**S&P Dow Jones  
Indices**

A Division of **S&P Global**

**Indices**

**Research & Insights**

**Exchange Relationships**

**Professional Resources**

**Governance**



[English](#) ▾ [About](#) ▾ ▾ [Discover more](#)  [Get S&P Global's offerings](#)

[English](#) ▾ [About](#) ▾ ▾ [Discover more about S&P Global's offerings](#)

## S&P Dow Jones Indices

A Division of **S&P Global**

## Indices

## Research & Insights

## Exchange Relationships

## Professional Resources

## Governance

[English](#) ▾ [About](#) ▾ ▾ [Discover more](#) [Register](#) [Log in](#) [Feedback](#) [Help](#) [Contact Us](#) [S&P Global's offerings](#)

**S&P Dow Jones  
Indices**

A Division of **S&P Global**

[Indices](#)

[Research & Insights](#)

[Exchange Relationships](#)

[Professional Resources](#)

[Governance](#)



[English](#) ▾ [About](#) ▾ ▾ [Discover more about S&P Global's offerings](#)

## S&P Dow Jones Indices

A Division of **S&P Global**

## Indices

## Research & Insights

## Exchange Relationships

## Professional Resources

## Governance



[English](#) ▾ [About](#) ▾ ▾ [Discover more](#) about S&P Global's offerings

## S&P Dow Jones Indices

A Division of **S&P Global**

## Indices

## Research & Insights

## Exchange Relationships

## Professional Resources

## Governance

[English](#) ▾ [About](#) ▾ ▾ [Discover more](#) at S&P Global's offerings

## S&P Dow Jones Indices

A Division of **S&P Global**

## Indices

## Research & Insights

## Exchange Relationships

## Professional Resources

## Governance

[English](#) ▾ [About](#) ▾ ▾ [Discover/ register](#) [Contact](#) [S&P Global's offerings](#)

## S&P Dow Jones Indices

A Division of **S&P Global**

## Indices

## Research & Insights

## Exchange Relationships

## Professional Resources

## Governance



About S&P Dow Jones Indices

Our Services

Media Center

Contact Us

Careers

Corporate Responsibility

History

Investor Relations

Leadership

